Some Methodological Issues in Economic Evaluation in Health Care

J.L. Severens

Some Methodological Issues in Economic Evaluation in Health Care

SOME METHODOLOGICAL ISSUES IN ECONOMIC EVALUATION IN HEALTH CARE

een wetenschappelijke proeve
op het gebied van de Medische Wetenschappen

Proefschrift

ter verkrijging van de graad van doctor aan de Katholieke Universiteit Nijmegen,
volgens besluit van het College van Decanen in het openbaar te verdedigen
op donderdag 2 december 1999 des namiddags om 3.30 uur precies

door

Johan Louis Severens

geboren op 9 juli 1963 te Voerendaal

Promotores

Prof. dr. P.F. de Vries Robbé

Prof. dr. F.F.H. Rutten (Erasmus Universiteit Rotterdam)

Co-promotor

Dr. G.J. van der Wilt

Manuscriptcommissie

Prof. dr. W.A.J. van Daal, voorzitter

Prof. dr. E.K.A. van Doorslaer (Erasmus Universiteit Rotterdam)

Prof. dr. P.P.M. Bossuyt (Universiteit van Amsterdam / AMC)

This thesis was supported by Astra Pharmaceutica BV, Nexstar Pharmaceuticals, Pharmacia & Upjohn BV, Janssen-Cilag BV, Pfizer BV, Sanofi Pasteur, SmithKline Beecham, and Siemens Nederland NV.

Some Methodological Issues in Economic Evaluation in Health Care

ISBN 90-9013028-4

Printed by: Ponsen & Looijen BV, Wageningen, the Netherlands

Aan mijn ouders

TABLE OF CONTENTS

CHAPTER 1
GENERAL INTRODUCTION

BACKGROUND

In 1997 in the Netherlands 8.6% of the gross domestic product was spent on health care (Statistics Netherlands, 1999) and this indicates the relevancy of the economic aspects of health care. As in many countries, in our country containment of health care expenditures is an important political topic. Thus, nowadays, efficacy can not be the only criterion on which basis choices to implement medical technologies can be made. Efficiency or cost-effectiveness of medical technologies, being the relation between outcome and input (resources, costs), is considered when choices are being made on the macro level (e.g. the Ministry of Health) as well as on meso- and micro-level (e.g. management of health care institutions and the practising health care workers)(Luce & Brown, 1995). The decisions have to be made about the availability and use of both therapeutic interventions and diagnostic technologies. By performing economic evaluations, health economic researchers intend to give decision-makers information about the relative efficiency of medical technologies by comparing alternative courses of action for consideration (Drummond *et al.*, 1997). In the past decades an increasing trend in the number of reported economic evaluations in the medical literature has been identified (Elixhauser, 1993a; Elixhauser *et al.*, 1993b).

The methodology of economic evaluation is still evolving. This evolution can be illustrated not only by the rising number of economic evaluations, but also by the increasing number of health economic journals which (partly) concentrate on methodological issues of performing economic evaluations. Discussion about these methodological issues is still going on and will probably continue for years to come. However, methodological principles of the economic evaluations that are currently executed influence the study results. This is one reason why several countries have implemented guidelines for performing economic evaluations either for specific subjects or economic evaluations in general (Access and Financing Division, 1998; Canadian Coordinating Office for Health Technology Assessment, 1996; Canadian Coordinating Office for Health Technology Assessment, 1997; Langley, 1996; National Health Insurance Board, 1999; Rutten *et al.*, 1993). The existing guidelines can be of help when performing economic evaluations of medical technologies. However, despite the guidelines, comparability of the results of economic evaluation can be difficult. When faced with the difficult task of assessing study results a critical appraisal checklist can be used which was published by Drummond *et al.* in 1987 (Drummond *et al.*, 1987) and revised in 1997 (Drummond *et al.*, 1997). This 10-item checklist may help to identify the key elements and assess the characteristics of an economic evaluation on which the study results are based. The items of this checklist can be divided into methodological and non-methodological issues, the latter being issues such as the way the research question was formulated, the timing of the study, and how the study findings were reported. The methodological issues mentioned in the checklist are the issues of the competing alternative,

relevant costs and consequences, accurate measurement of costs and consequences, credible valuing of costs and consequences, differential timing, incremental analysis, and the issue of uncertainty.

AIM OF THE THESIS

In this thesis several methodological issues for the economic evaluation of health care technologies will be discussed. This thesis addresses research questions on the following issues in economic evaluation:

- The choice of the competing alternative;
- The relevant costs and consequences;
- Accurate measurement of costs and consequences;
- Credible valuing of costs and consequences; and
- The uncertainty of the results.

Finaly, regarding these issues, a comparison is made of the different guidelines for performing economic evaluations.

OUTLINE OF THE THESIS

In Chapter 2, the issue of the choice of the competing alternative is described. The choice of the alternative to which the intervention under study is compared to is essential for the findings of a study. For decision-makers, the comparator should be relevant in the sense that the study should reflect the actual decision at stake. Chapter 2.1 concentrates on the competing alternative in the situation where a diagnostic technology is being evaluated. In the case of economic evaluation of diagnostic technologies, there are various categories of alternatives for comparison: test vs. no-test, test A vs. test B, and so on. Diagnostic facilities are hardly ever used solitary and in this case the evaluation of a specific test should be performed in the context of alternative tests, or more specifically, in the context of the sequence of tests. When more than one diagnostic test is considered, decisions have to be made not only about which diagnostic tests to perform, but also about the sequence of testing. A method is described to evaluate the use of a diagnostic test in the context of the use of alternative tests. The aim of this method was to explore a model for optimising the sequence of diagnostic tests based on the principles of a cost-minimisation analysis. Thus, an emphasis has been put on efficient sequence of testing without losing any diagnostic information. First, the principles of the model are described, followed by a description of two applications of the model: diagnosis of Helicobacter Pylori and benign prostatic hyperplasia. Chapter 2.2 focuses on the issue of the competing alternative in the situation of comparing therapeutic alternatives. As with evaluating diagnostic technologies, the question of a therapeutic comparator is essential to the findings of a study. For instance, for licensing purposes most pharmaceutical

trials are based on a placebo comparison while for policy decisions, a comparison with normal care or the best alternative would be relevant. Modelling, which involves applying mathematical techniques to synthesis available information concerning a therapeutic intervention, can be used to extend comparators. Besides, modelling can be useful to extend trail results in the situation that effectiveness has been studied but costs were not subject to analyses. In this chapter a decision analytic cost–effectiveness comparison is made of different empirical treatment strategies of invasive fungal infection (IFI) in patients with haematological malignancy. In the first strategy, amphotericin B desoxycholate (DC-Amb) was given as first line empirical treatment of IFI and was changed to liposomal amphotericin B (L-Amb) in case of a treatment failure or nephrotoxicity (DC/L-Amb strategy). In the other strategy, L-Amb was used for empirical first line therapy of IFI (L-Amb strategy). As far as we know the costs and effectiveness of these strategies have never been studied prospectively or retrospectively.

The choice of the costs and consequences that are to be analysed in an economic evaluation depends on their relevancy. Two aspects play an important role when choosing costs and consequences. First, the time horizon used should extend far enough in the future to capture the major health and economic outcomes. Second, the viewpoint or perspective of a study can influence the choice of the cost calculation method. These aspects are related to the issue that is described in Chapter 3: the relevant costs and consequences in an economic evaluation. In Chapter 3.1 the importance of determining the time horizon is examined. Determining the time horizon of a study is especially relevant when evaluating diagnostic tests, because this might influence the study findings. In principle, the process of diagnosis focuses on reducing uncertainty about the presence of a disease of a person. Hence, the number of cases detected is an effectiveness parameter that is used regularly in this type of research. However, this parameter can be regarded as an intermediate outcome measure in the sense that it does not reflect the actual health outcome as a result of diagnosis and possible treatment of a patient. In case it is not possible to prospectively measure health outcome because of the limited time horizon of the prospective part of the study, modelling can be a solution to extend the time horizon of the analysis. A decision analytic model is used to compare alternative strategies for diagnosis and treatment of invasive aspergillosis. Chapter 3.2 illustrates the relevancy of the methods used to determine costs involved in medical technologies. The methods to analyse costs of health care technologies are highly dependent on the perspective of the economic evaluation. This is demonstrated by an analysis of the costs related to cochlear implants for children where after an explicit comparison with results of cost analyses of other studies is made.

Accurate measurement in appropriate physical units of costs and consequences is discussed in Chapter 4. Incorporating productivity costs in an economic evaluation can have a substantial influence on the study findings; however, the methods used to measure

productivity costs are still being developed[1]. Retrospective measurement of absence from work as a basis for calculating productivity costs is often used. Different recall periods up to twelve months can be found in the literature. Precision and accuracy measuring absence from work retrospectively is studied using both questionnaire data and prospectively registered data.

In Chapter 5, the issue of credible valuing of costs and consequences is dealt with. Chapter 5.1 concentrates on valuing productivity costs. As stated before, incorporating these costs can have a substantial influence on the study findings. However, valuing each day of absence from work merely reflects potential productivity costs instead of real productivity costs. The impact of considering compensating mechanisms for not being able to work when calculating these type of costs is shown using diagnosis and treatment of patients having dyspeptic complaints.

Valuing consequences of a diagnostic technology can be done in several ways. Often the consequences are reflected in measures such as the number of patients diagnosed accurately or the impact of the physician's decision. However, the principle reason to perform diagnostic tests is to gain information and the questions arises if the diagnostic information as such can be valued. Willingness to pay analysis was used as a measure of outcome to reflect value of the non-decisional diagnostic information to persons at risk for histoplasma capsulatum. In Chapter 5.2 several hypotheses related to the value of non-decisional diagnostic information are presented and studied in order to investigate the construct validity of willingness to pay measurement.

The issue of the uncertainty of the results of an economic evaluation is dealt with in Chapter 6. The summary outcome of an economic evaluation in general is an incremental cost-effectiveness ratio, which is calculated by dividing difference in costs between alternatives by difference in consequence or effect. This ratio reflects the investment that is necessary to gain one unit of effect. However, this ratio is a point estimate that does not give insight into the uncertainty of the study findings. In this chapter a comparison is made between different statistical methods to study the uncertainty of this ratio using data from a prospective randomised trial.

In Chapter 7 the findings of the previous chapters are briefly summarised and discussed in the light of the aims of this thesis. A comparison is made of the different guidelines for the economic evaluation of health care technologies that exist in several countries.

---

1. The term 'productivity costs' is recommended by the Panel on Cost Effectiveness in Health and Medicine, appointed by the US Public Health Service, as an alternative term for indirect non-medical costs. The term reflects the costs of lost productivity due to absence from work (Luce *et al.*, 1996).

CHAPTER 2
THE ISSUE OF THE CHOICE OF THE COMPETING ALTERNATIVE

CHAPTER 2.1
THE ISSUE OF THE CHOICE OF THE COMPETING ALTERNATIVE: DIAGNOSTIC TEST SEQUENCE

Based on Severens JL, Vries Robbé PF de & Verbeek ALM (1999). Optimising diagnostic test sequences: the probability modifying plot. *Methods of Information in Medicine* 38: 50-55,

and Severens JL, Sonke GS, Laheij RJF, Verbeek ALM & Vries Robbé PF de. Efficient diagnostic test sequence: applications of the probability modifying plot [submitted for publication].

INTRODUCTION

An important issue when performing an economic evaluation is the choice of the competing alternative to which the intervention under study is compared, because this can influence study findings (Berger, 1995). For decision-makers, the comparator should be relevant in a sense that the study should reflect the actual practical decision at stake. In the case of evaluating diagnostic facilities, there are various categories of alternatives for comparison: for instance test versus no test, one test versus another test, and test versus direct treatment. In a literature review it was shown that the test versus test comparison is performed most (Severens & van der Wilt, 1999b). However, diagnostic facilities are hardly ever used solitary and therefore the evaluation of a specific test should not only be performed in contrast to other tests but also, when relevant, in the context of the other tests. On the assumption that time delay is of no importance, sequential use of diagnostic tests is potentially more efficient than performing multiple tests simultaneously (Doubilet & Cain, 1985). Therefore, decisions have not only to be made about which diagnostic tests to perform, but also about the sequence of testing. In the former case, the marginal gain of each test in a sequence can be determined and knowledge from previously performed tests can be used to decide which, if any, test should be performed next. In case many factors such as test characteristics and costs of numerous diagnostics tests are to be included in diagnostic strategy considerations, the physician's ability to intuitively integrate the results seem to be reaching a limit (Henschke *et al.*, 1997). There is a need for a methodology for evaluation of different testing strategies.

Several methods can be used to evaluate the optimal sequence of diagnostic tests. Most methods which are designed to evaluate test sequences, however, are based on the principle of including tests with respect to their gain in certainty of disease status. The quest for diagnostic certainty about a specific disease can cause excessive testing, but complete certainty about the presence of disease is rarely achieved (Kassirer, 1989; Putterman & Ben-Chetrit, 1995). The choice of effective diagnostic tests is crucial, but should not only be based on test characteristics such as sensitivity and specificity; probability of disease and utilities of correct and incorrect disease classification should also be considered (Boyko, 1994). The decision problem of which diagnostic test to use and besides this, comparison of diagnostic strategies can be structured in decision trees (Kent *et al.*, 1995; Erkel *et al.*, 1996; Goldberg Kahn *et al.*, 1997; Severens *et al.*, 1997). However, unmanageable bushy decision trees result when all theoretically possible strategies of two or more tests are investigated instead of a predetermined strategy. Besides, incomplete and inconsistent decision trees might result from this exercise. Decision tables, which are a simplification of bushy decision trees, can be used to overcome these problems (Glasziou & Hilden, 1986). Despite this advantage, decision tables are under-utilised, for which the reason is not clear (Glasziou, 1994).

In this chapter a method is described to evaluate the use of a diagnostic test in the context of the use of alternative tests. The aim of this method was to explore a model for optimising the sequence of diagnostic tests based on efficiency criteria, thus using the principles of cost-minimisation analysis. This means that an emphasis has been put on efficient sequence of testing without losing diagnostic information relevant for the treatment decision.

METHODS

Data from a study on urinary tract infection were used to develop the model. The data concerned the results of eleven diagnostic tests which examined the urine from a group of 550 patients who were suspected of having urinary tract infection. For reasons of simplicity, all test results were transformed into dichotomous outcomes. Besides these tests, a urine culture was performed which was used as a gold standard on the basis of which the pre-test probability of disease in this population was determined to be 0.145 (80/550). Disadvantages of the urinary culture are that it is expensive and that it takes several days before the test results are available (Lachs *et al.*, 1992). Therefore, other, faster tests can be used to determine whether treatment is indicated. Analysis of the data showed that only three of out of the eleven tests made a significant contribution to predicting the presence of urinary tract infection, defined as having a positive urinary culture. These tests will be called T_1, T_2 and T_3. In summary, the data of T_1, T_2 and T_3 and the urinary culture were selected to develop the test sequence model[2].

Defining the treatment threshold

The concept of a treatment threshold reflects the fact that decisions about treating a patient are usually made under uncertainty. If the probability of a person having the disease is low, treatment is not indicated; if the probability of disease is high, treatment is indicated. The estimation of the probability of disease at which one is indifferent as to giving treatment and withholding treatment is called the treatment threshold probability. This treatment threshold probability can be estimated intuitively or calculated from the net benefits and cost of treatment (Sox *et al.*, 1988). This threshold is a result of weighing the following situations (Pauker & Kassirer, 1980; Kassirer, 1989):

- a healthy person, not being treated,
- a person having the disease, not being treated,
- the deterioration in health status by treating a healthy person, and

2. Unpublished; the data on urinary tract infection are solely used as an illustration for our model. The three test included in the model are Gram preparation, which is a microscopic test to determine the presence of bacteria; Kova microscopic test to determine the number of bacteria; Miditron urine dipstick to determine the presence of nitrite.

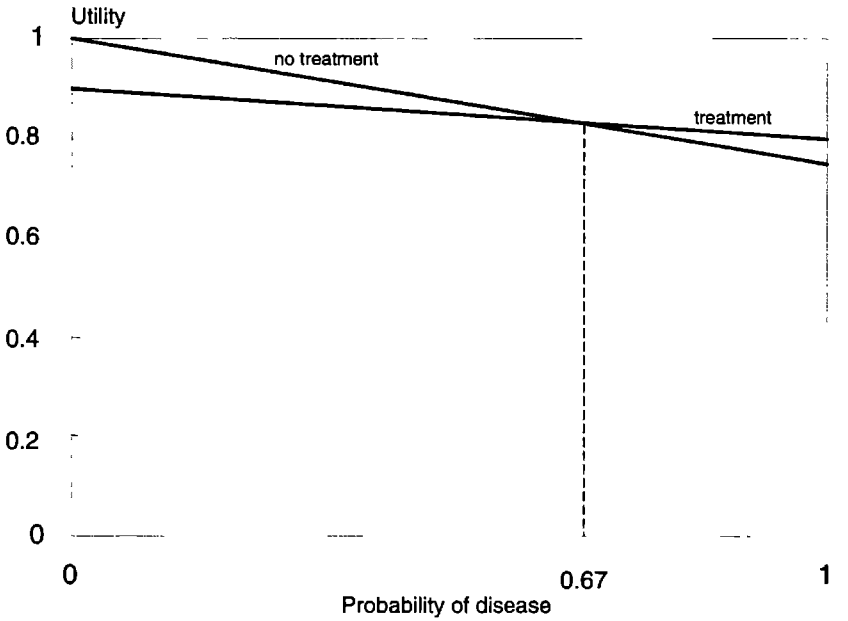


Figure 2.1 The graph visualises the relationship between the probability of disease and the utility of possible treatment or withholding treatment decisions. The intersecting lines indicate the expected utility of withholding treatment and the expected utility of treatment. The treatment threshold probability of having the disease is determined to be 0.67.

– the gains in health status by treating a person having the disease.

Consequently, determining the treatment threshold is possible when the four possible situations are valued, which may occur after having made a decision about treatment or no treatment.

In our example, the treatment threshold is estimated for the decision about treating or not treating a person suspected of having urinary tract infection. A specific health state can be given a utility, which can be defined as a value between 0 and 1 which reflects the relative value of one health state to another. Suppose a healthy person (a person without urinary tract infection) is given the highest possible utility 1. A person who has urinary tract infection and who is not treated will have a lower utility because of the burden of the disease, for instance 0.75. In Figure 2.1 these two levels of utility are indicated on the left and right ordinate y. The probability of disease is indicated on the x-axis. The line which is drawn between the points is called the expected utility line of withholding treatment and indicates which utility can be expected when treatment is being withheld, given a certain probability of disease.

In this example, this line is defined by $y = 1 - 0.25x$, where y is the expected utility and x is the probability of disease. An expected utility line of treatment can be drawn between the next two points. If a healthy person is treated for urinary tract infection, the

burden of treatment (and suspicion of having the disease) will decrease the utility, for instance, the utility will become 0 9 A sick person who is treated will gain utility from being treated, but will not have a perfect utility of 1 because of the burden of treatment, for instance 0 8 In this example, the expected utility line of treatment is indicated by $y = 09 - 01x$ The intersection between the two lines represents the treatment threshold probability At this point there is indifference between expected utility of treatment and withholding treatment, hence at $1 - 025x = 09 - 01x$ This equation results in a treatment threshold probability of 0 67 As can be seen from the graph, at a lower probability of disease the expected utility of withholding treatment is higher than the expected utility of treatment and, therefore, withholding treatment is the best thing to do A probability of disease which is higher than the treatment threshold probability indicates treatment as the choice with the highest utility For our example, the utilities on which the treatment threshold is based are estimations Depending on the defined utility lines, the treatment threshold can be sensitive to changes in utilities In summary, the treatment threshold probability approach reflects the utilities which are given to both situations of misclassification healthy persons misclassified as being sick, and sick persons misclassified as being healthy

Table 2 1 Post-test probabilities of test combinations in relation to the treatment threshold probability

T_1	T_2	T_3	D^+	D	post-test probability	threshold
	+	+ (32)	31	1	$P(D^+ \mid T_1^+, T_2^+, T_3^+) = 0\ 97$	>
+	(36)	- (4)	3	1	$P(D^+ \mid T_1^+, T_2^+, T_3) = 0\ 75$	>
(61)	-	+ (12)	10	2	$P(D^+ \mid T_1^+, T_2, T_3^+) = 0\ 83$	>
	(25)	- (13)	6	7	$P(D^+ \mid T_1^+, T_2, T_3) = 0\ 46$	<
	+	+ (35)	9	26	$P(D^+ \mid T_1, T_2^+, T_3^+) = 0\ 26$	<
-	(101)	- (66)	6	60	$P(D^+ \mid T_1, T_2^+, T_3) = 0\ 09$	<
(489)	-	+ (44)	5	39	$P(D^+ \mid T_1, T_2, T_3^+) = 0\ 11$	<
	(388)	- (344)	10	334	$P(D^+ \mid T_1, T_2, T_3) = 0\ 03$	<

> post-test probability higher than treatment threshold probability
< post-test probability lower than treatment threshold probability
( ) number of patients between parentheses

Sequential testing

When more than one test is considered to determine the presence of a disease, a decision table is made which contains an overview of the possible combinations of test results (Table 2.1). Given three dichotomous tests in our illustration, eight combinations of test results are possible. For each of these combinations the post-tests probability of having urinary tract infection is determined on the basis of the gold standard. Furthermore, for each test result combination, the post-tests probability is compared to the treatment threshold. The final column of this table indicates whether the test result combination leads to a probability higher or lower than the treatment threshold. In the former case treatment is indicated; in the latter no treatment is indicated. As Table 2.1 shows, the combinations of test results (T_1^+,T_2^+,T_3^+), (T_1^+,T_2^+,T_3^-) and (T_1^+,T_2^-,T_3^+) indicate treatment. The other combinations of test results indicate no treatment because the post-tests probability is lower than the treatment threshold probability. A decision table gives a clear overview of the possible combinations of test results, but it does not indicate whether applying a diagnostic test is useful at all.

Test thresholds

Applying a diagnostic test or a combination of diagnostic tests is only considered to be useful if it can change the probability of disease from being lower to being higher than the threshold probability (or the other way around). Only if this is the case, a test result will influence the therapeutic decision. Successively, the range of pre-test probabilities can be determined for which testing is useful. In this way, the pre-test probability of having the disease is part of the model. Two threshold probabilities have to be calculated, the no-treat/test threshold probability and the test/treatment threshold probability (Sox *et al.*, 1988). Between these threshold probabilities testing is indicated. These threshold probabilities are calculated on the basis of the defined treatment threshold probability and the likelihood ratios of positive combinations of test results and negative combinations of test results[3]. In our example, three combinations of test results, (T_1^+,T_2^+,T_3^+) or (T_1^+,T_2^+,T_3^-) or (T_1^+,T_2^-,T_3^+), indicate presence of disease. As can be summarised from Table 2.1, 44 out of the 80 diseased people had one of the above mentioned positive combinations of test results, which leads to a sensitivity of 0.55. The specificity is calculated to be 0.99 (466/470). The sensitivity and specificity lead to a likelihood ratio of negative combinations of test results of 0.45 (thus, a negative result is 0.45 more likely in patients than in non-patients). This likelihood ratio and the treatment threshold probability of 0.67 are used to calculate the test/treatment threshold probability, which is 0.82. Similarly, the likelihood ratio of the positive combinations of test results is calculated to be 55 (thus, a positive result is 55 more likely in patients than in non-patients). Again, when the treatment threshold probability and this likelihood ratio are considered, the no-treat/test

3. For reasons of simplicity, a combination of test results which indicates the presence of disease is called positive. A combination of test results which indicates the absence of disease is called negative.

threshold probability is found to be 0.04. In conclusion, two threshold probabilities have to be calculated between which applying diagnostic tests is useful, regarding the treatment/no-treatment decision. Given the pre-test probability of 0.145 in our population, testing is useful.

The probability modifying plot

Once testing is determined to be useful, the probability modifying plot can be used as an instrument to visualise and examine the efficiency question about the sequence of tests. With our example on urinary tract infection, arbitrarily starting with T_1 as a first test, the probability of disease given a positive T_1 is 0.82 and the probability of disease after a negative T_1 is 0.11. The change from the pre-test probability to the probabilities after using T_1 as the first test can be seen in a probability modifying plot (Figure 2.2). The y-axis indicates the probability of having the disease, and the x-axis indicates the different steps in the diagnostic process: the stage after a first test, the stage after a first and second test, and the stage after all three tests. The horizontal line indicates the treatment threshold probability at 0.67. The points above each test mark on the x-axis indicate the probability after a certain test result or combination of test results. At each point in the plot, the specific test being used at a certain stage is mentioned. Because of the dichotomous tests, from each point in the plot two lines continue to the next test mark, one line ascending and one line descending. An ascending line indicates that a positive test will result in a higher probability of disease; a descending line indicates a lower probability after a negative test result. After a specific result from T_1, again an ascending and a descending line continue as a result of performing T_2. This diversion from each point in the plot continues until the eight possible post-test probabilities as mentioned in Table 2.1 are reached. These post-test probabilities are indicated on the right y-axis.

Efficiency of testing

To incorporate efficiency criteria in the probability modifying model, the principle of determining the usefulness of a test or a series of tests can be used. As can be seen from the probability modifying plot, after T_1^- the results of T_2 and T_3 being either T_2^+ and T_3^+, T_2^- and T_3^+, T_2^+ and T_3^-, or T_2^- and T_3^-, do not lead to a probability that lies above the treatment threshold probability. Therefore, once T_1 has a negative test result, further testing is not useful considering the treatment versus withholding treatment decision (Fendrick *et al.*, 1995). The situation after a positive result of T_1 is completely different. A negative result from T_2 (after T_1^+) can still lead to a probability which is higher or lower than the threshold. Therefore, all persons with T_1^+ should be subjected to T_2. After T_1^+ and T_2^+, the result of T_3 (either positive or negative) no longer influences the decision about treatment and is therefore considered to be redundant for this subgroup of patients. Although the persons with T_1^+ and T_2^- have a probability which is lower than the threshold, this situation is not yet certain given the fact that the result of a third test critically influences the probability of having the disease (0.46

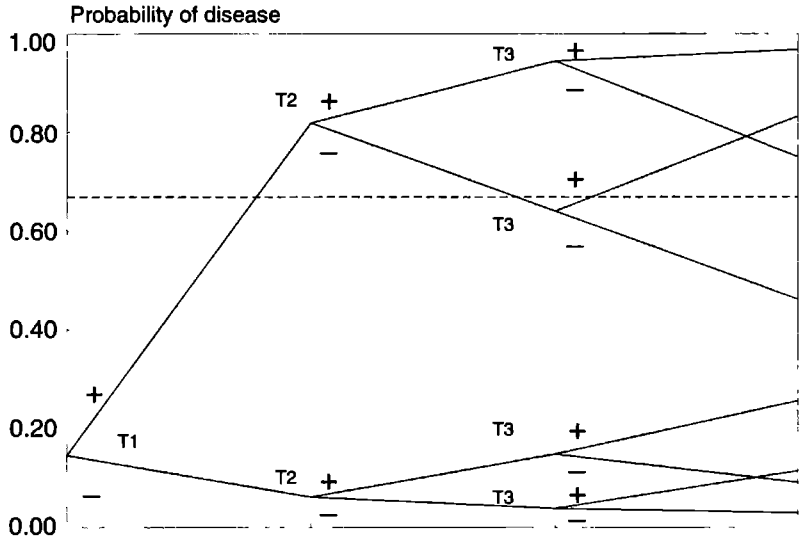


Figure 2.2 The probability modifying plot of the test sequence T_1, T_2, T_3 shows the treatment threshold probability and the changes in the probability of having the disease after specific test results. The plot indicates in which case further testing is redundant. For instance, when T_1 is negative, the results of the use of T_2 and T_3 no longer influence the treatment/no-treatment decision.

versus 0.83). With regard to the test sequence T_1,T_2,T_3, the result of the third test T_3 will be of importance considering the treatment decision for only those persons who already had test results T_1^+ and T_2^-. In conclusion, from the probability modifying plot it is clear when further testing is redundant.

With the help of the plot, the total number of tests needed or the total cost of testing can be determined for a specific test strategy. For each subgroup of patients it is possible to determine for a given sequence whether one, two, or three tests are necessary to decide between treatment or withholding treatment. Once the number of tests needed is determined, the cost of diagnostic tests involved can easily be included in the probability modifying model. With regard to our example of urinary tract infection, the prices of the tests are as follows: T_1: (Gram preparation) Dfl. 25.50; T_2: (Kova bacteria) Dfl. 1.70; T_3: (Miditron nitrite) Dfl. 6.80[4]. To decide which test sequence is most efficient, regarding either the criterion number of tests needed or the total cost of testing, all possible sequences must be evaluated.

4. One US dollar is approximately 1.72 Dutch guilder (Dfl.)

RESULTS

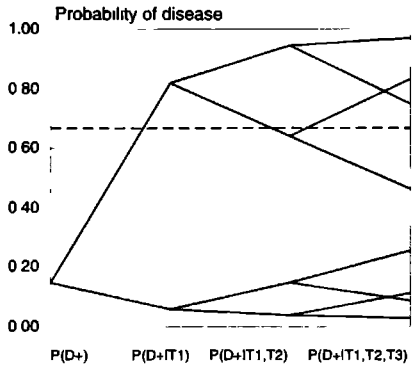
The probability modifying model is used to explore the efficiency of alternative sequences of tests. In the first probability modifying plot, the test sequence T_1, T_2, T_3 is given; however, there are six possible sequences of the three tests which all lead to the same final post-test probabilities. These six sequence strategies are shown in Figure 2.3. As can be seen from the different probability modifying plots in this figure, the incremental increase or decrease of the probability differs for each sequence and is dependent on the next test in a row. The question remains which sequence is the most efficient. Consider the original sequence T_1, T_2, T_3 in plot A. On the basis of this plot, the total number of tests needed to answer the treatment/no-treatment question is presented in Table 2.2.

To decide which test sequence is most efficient, all possible sequences must be evaluated as has been done with the sequence T_1, T_2, T_3. The results of the calculations for all six test sequences are presented in Table 2.3. As can be seen from this table, the test sequence T_1, T_3, T_2 gives the lowest number of tests needed to determine whether a person of this specific population has a disease probability that indicates treatment or withholding treatment. This is the most efficient sequence, considering the number of tests as a criterion. Incorporating the cost of testing gives the results as shown in the last column of Table 2.3. Despite the lowest number of tests, strategy 2 does not have the lowest total cost for testing. Sequences 3 and 5 have the least cost and are therefore the most efficient test sequences if cost is to be the criterion. The combination of both T_2 and T_3 selects a number of persons for whom the expensive T_1 is no longer necessary. It is not possible to make a distinction between strategies 3 and 5 on the basis of the cost criterion alone.

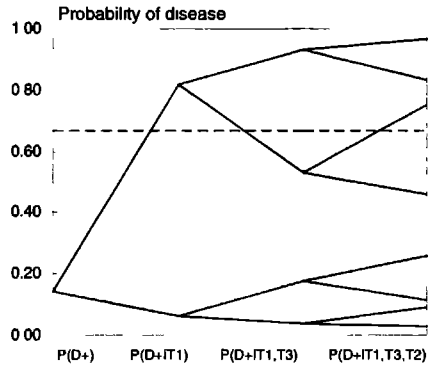
Table 2.2 Calculation of the total number of tests needed, sequence T_1 T_2 T_3.

T_1	T_2	T_3	number of tests
+ (61)	+ (36)	no further testing	(36) * 2
	- (25)	+ (12)	(12) * 3
		- (13)	(13) * 3
- (489)	no further testing		(489) * 1
total number of tests			636

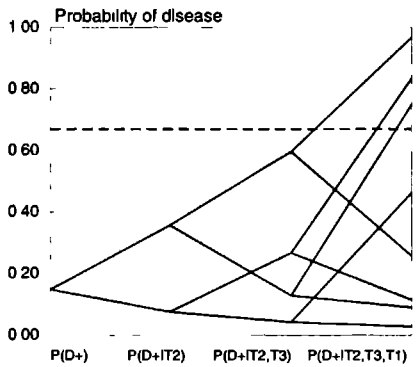
( ): number of patients between parentheses



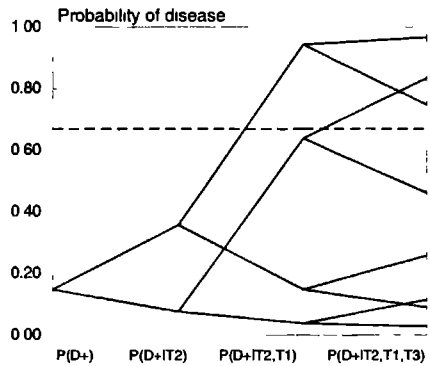
Plot A



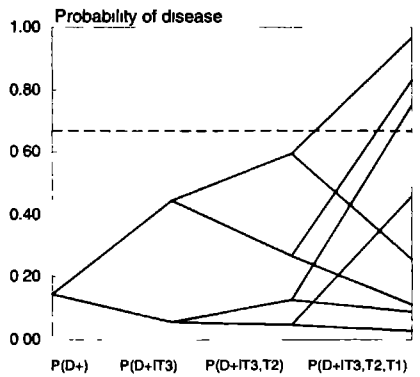
Plot B



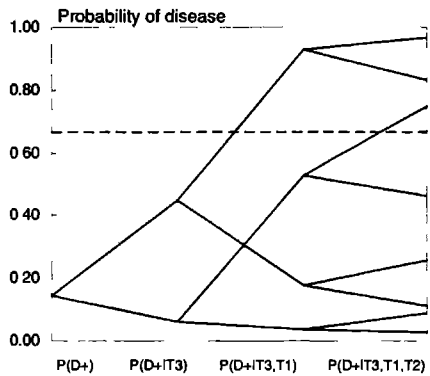
Plot C



Plot D



Plot E



Plot F

Figure 2.3 In the situation of three dichotomous tests, six different sequences of test strategies exist. The incremental value of information of each test can be noted by the slope of the lines in the plots. These plots are the basis for calculating the number of tests needed and the total costs of testing.

Table 2.3 Number of tests needed and cost of testing, given all possible test sequences.

sequence	first test	second test	third test	number of tests	cost of testing (Dfl.)
$T_1 T_2 T_3$	550 * T_1	61 * T_2	25 * T_3	636	14,299
$T_1 T_3 T_2$	550 * T_1	61 * T_3	17 * T_2	628	14,469
$T_2 T_3 T_1$	550 * T_2	550 * T_3	193 * T_1	1,293	9,597
$T_2 T_1 T_3$	550 * T_2	550 * T_1	25 * T_3	1,125	15,130
$T_3 T_2 T_1$	550 * T_3	550 * T_2	193 * T_1	1,293	9,597
$T_3 T_1 T_2$	550 * T_3	550 * T_1	17 * T_2	1,117	17,794

APPLICATIONS

We described a model to optimise the sequence of diagnostic tests on the basis of efficiency, illustrated with data on urinary tract infection. Extensive validation of the test results according to this specific disease was beyond the scope of this description; no definite conclusions can be drawn about the specific test sequence for urinary tract infection. However, we applied the model using two databases on diagnostic tests: diagnosis of Helicobacter Pylori and Benign Prostatic Hyperplasia.

Diagnosis of Helicobacter Pylori

Description of disease and diagnosis

Helicobacter Pylori is one of the most common chronic infections in humans. The infection may lead to substantial morbidity, and, in the long run, it may cause peptic ulcer or gastric cancer, thus mortality can be attributed to this phenomenon. However, widespread non-invasive testing, such as a breath test and serology, followed by treatment of non-symptomatic infected persons is beyond issues of feasibility and financial consideration. Besides this, information on possible unwanted side effects of this approach is lacking (Rabeneck & Graham, 1997). Therefore, whether a person is infected with Helicobacter Pylori only becomes relevant in persons with dyspepsia, which can be described as episodic or persistent upper abdominal pain or discomfort that is thought by the physician to arise in the upper gastrointestinal tract. Only the invasive upper endoscopy will reveal the cause of these complaints and is therefore the most widely used diagnostic tool (de Boer, 1997). Endoscopy

results in diagnosis of erosive esophagitis, peptic ulcer disease and gastric cancer only in one third of the times (Rabeneck & Graham, 1997). The question arises what to do with dyspeptic patients with no detectable organic disease. During endoscopy, multiple mucosal biopsy specimens can be taken from the stomach. To determine the presence of Helicobacter Pylori in the biopsies several tests can be used. Histologic examination is performed by a pathologist on several cuts of the biopsies and is based on the identification of micro-organisms with appropriate morphology, location and staining characteristics. Bacterial culture of the biopsies is based on the principle that Helicobacter Pylori bacteria can grow on media in the laboratory (Barthel & Dale Everett, 1990). Rapid urease tests are based on demonstrating urease activity in the biopsy specimen and these test can easily be performed in the endoscopy suite. Besides these, polymerase chain reaction and phase contrast microscopy can be used but are left out of our analyses because they presently have no significant clinical application (de Boer, 1997).

Patient data

The patient data which were used for our probability modifying model concerned a study population of consecutive patients undergoing routine endoscopy at a general hospital in the Netherlands. The patients included in the study all had dyspeptic complaints and had been sent to the outpatient clinic by a specialist or a general practitioner for open-access endoscopy. Endoscopy was performed and biopsies were taken. Two biopsies were used for bacterial culture, histologic examination was performed on another two biopsies, and one biopsy was used for a CLO-test, a specific make of a rapid urease test, which was analysed after 24 hours. The materials of 869 endoscopic procedures regarding 627 patients were used in our analyses. An extensive description of the data can be found elsewhere (Laheij *et al.*, submitted).

Application of the model

As mentioned earlier, the treatment threshold probability can be estimated intuitively. In the case of Helicobacter Pylori this can be done by weighing the situations of treating a Helicobacter Pylori positive person, not treating a Helicobacter Pylori positive person, and treating a Helicobacter Pylori negative person. In case of Helicobacter Pylori it can be argued that the health status of a healthy person without Helicobacter Pylori and a person who is successfully treated for Helicobacter Pylori is equal. In case the relatively low level in health state of a Helicobacter Pylori positive person not being treated and the burden of treatment of a Helicobacter Pylori negative person would estimated to be equal, this would result, for instance, in $y = 1 - 0.1x$ for withholding treatment and $y = 0.9 + 0.1x$ for treatment. The treatment threshold would therefore be 0.50 (Figure 2.4). For the situation that a Helicobacter Pylori positive person is not being treated and a Helicobacter Pylori negative person is being treated it can be motivated that the first situation is worse because dyspeptic symptoms have a negative impact on health status whereas falsely being treated only have a slight impact. Presume a proportion of 1 to 2 for both situations of misclassification; this can be expressed as $y = 1 - 0.2x$ for withholding treatment and, given $y = 0.9 + 0.1x$ for treatment, the threshold

would become 0.33 (Figure 2.4). Thus, in this situation, the absolute utilities of the health states are not relevant; the relation or proportion between both situations of misclassification is decisive when determining the threshold (Phelps, 1997).

Table 2.4 reflects the decision table based on our data. Again, given three dichotomous tests, eight combinations of test results are possible. The post-test probability for each of these combinations in the Helicobacter Pylori diagnosis case is determined on the basis of a gold standard which was ascertained by a least square method (Laheij *et al.*, submitted). Four of the eight possible test result combinations indicate presence of Helicobacter Pylori, in fact all combinations in which at least two tests are positive. As can be summarised from this table, sensitivity of these test result combinations is 0.997 (342/343), and specificity 0.988 (520/526). The likelihood ratio of positive test combinations and negative test combinations are therefore 83.1 and 0.003 respectively. Given the defined treatment threshold and these likelihood ratios, calculation of the test-thresholds results in a no-treat/test threshold probability of 0.012 and a test/treatment threshold probability of 0.997. Regarding the pre test probability of disease in our study population of 0.40 it can be concluded that testing is indicated.

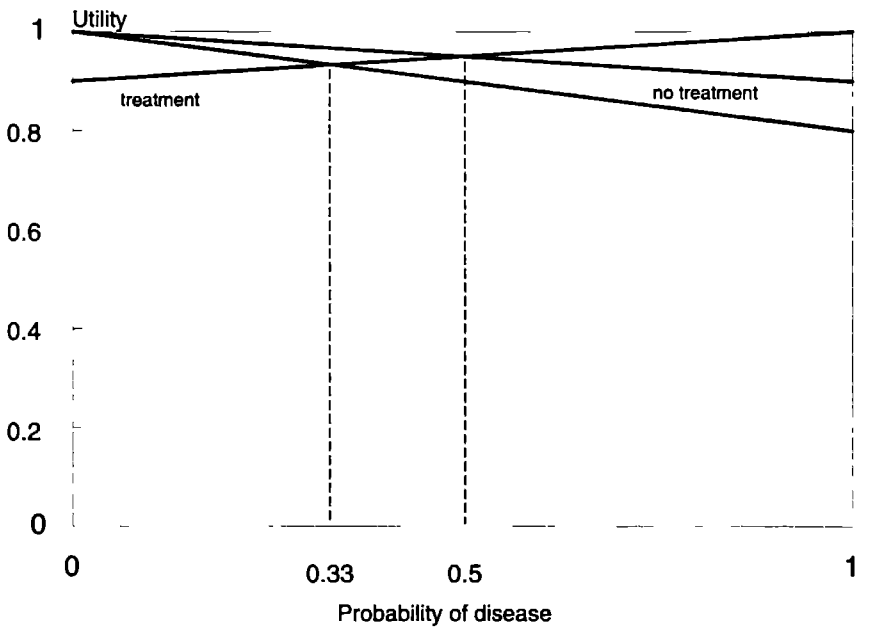


Figure 2.4 Plot showing the treatment threshold probabilities for different proportions between two situations of misclassification: the decline of utility because of withholding treatment to a diseased person on the one hand and the decline of utility because of treating a healthy person on the other hand.

Table 2 4 Post-test probabilities of test combinations of Helicobacter Pylori tests in relation to the treatment threshold probability

C	H	U	D^+	D	post-test probability	threshold
	+	+ (272)	272	0	$P(D^+ \mid C^+, H^+, U^+) = 1\ 00$	>
+	(287)	- (15)	15	0	$P(D^+ \mid C^+, H^+, U) = 1\ 00$	>
(337)	-	+ (30)	29	1	$P(D^+ \mid C^+, H, U^+) = 0\ 97$	>
	(50)	- (20)	2	18	$P(D^+ \mid C^+, H, U) = 0\ 10$	<
	+	+ (26)	26	0	$P(D^+ \mid C, H^+, U^+) = 1\ 00$	>
-	(38)	- (12)	1	11	$P(D^+ \mid C, H^+, U) = 0\ 08$	<
(532)	-	+ (19)	3	16	$P(D^+ \mid C, H, U^+) = 0\ 16$	<
	(494)	- (475)	0	475	$P(D^+ \mid C, H, U) = 0\ 00$	<

C bacterial culture, H histologic examination, U rapid urease test
> post-test probability higher than treatment threshold probability
< post-test probability lower than treatment threshold probability
( ) number of patients between parentheses

Again, once testing is determined to be useful, the probability modifying plots can be drawn (Figure 2 5) Because culture, histology and urease test all have a rather high sensitivity and specificity the six plots look more or less the same This is confirmed when the calculations for the number of tests needed for the various test sequences are compared (Table 2 5) Only in a relatively small number of cases a third test is necessary to finally determine the Helicobacter Pylori infection status of a patient, related to the treatment threshold Using the total number of tests needed as an efficiency criterion, a straightforward preference can hardly be made However, when the costs for the different test are incorporated in the calculations, it seems to be clear that the sequence in which histologic examination is performed last and only in patients for which culture and the urease test do contradict, is the sequence which is most efficient Regarding the costs of testing which can be expected for each patient these sequences result in Dfl 48 The sequences in which either culture or urease is performed last result in Dfl 96 and Dfl 115 per patient, respectively

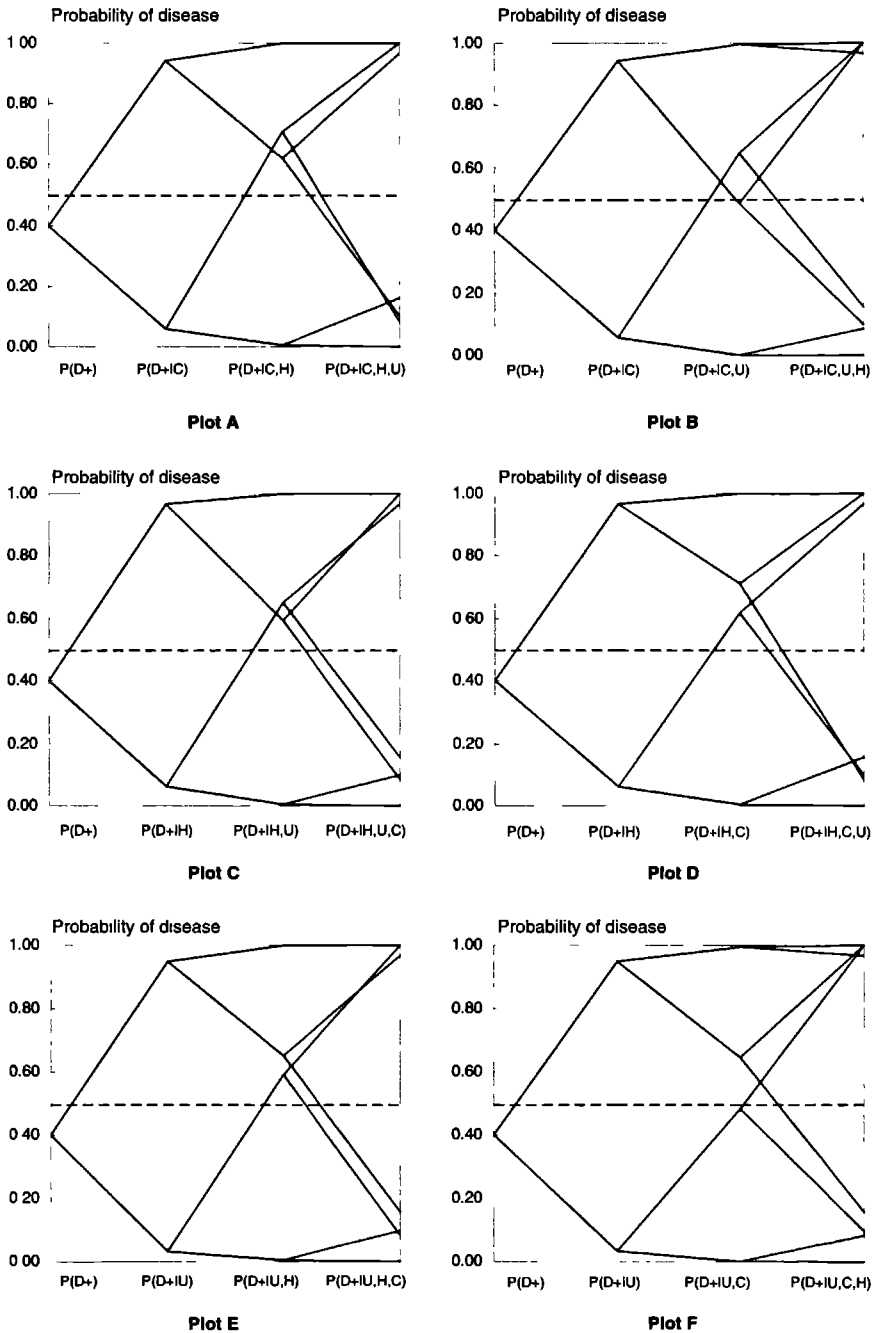


Figure 2.5 Overview of 6 plots showing the different test sequences regarding bacterial culture (C), histologic examination (H), and the urease test (U) for determining the presence of Helicobacter Pylori.

Table 2.5 Number of tests needed and cost of testing, given all possible test sequences for diagnosis of Helicobacter Pylori.

sequence	first test	second test	third test	number of tests	cost of testing[5](Dfl.)
C H U	869 * C	869 * H	88 * U	1,826	99,772
C U H	869 * C	869 * U	80 * H	1,818	41,937
H U C	869 * H	869 * U	76 * C	1,814	83,316
H C U	869 * H	869 * C	88 * U	1,826	99,772
U H C	869 * U	869 * H	76 * C	1,814	83,316
U C H	869 * U	869 * C	80 * H	1,818	41,937

C: bacterial culture; H: histologic examination; U: rapid urease test.

Diagnosis of Benign Prostatic Hyperplasia

Description of disease

Benign Prostatic Hyperplasia (BPH) is a common disease in elderly men. BPH is a histologic condition which may give rise to a benign enlargement of the prostate. The enlargement of the prostate gland can cause progressive blockage of urinary flow. Thus, four properties are related to the disease process of BPH: 1) histologic BPH 2) macroscopic benign enlargement of the prostate gland (BPE), 3) obstruction of the bladder outlet (BOO) and 4) lower urinary tract symptoms (LUTS), such as diminution in the calibre and force of the urinary stream, hesitancy initiating voiding, inability the terminate micturition abruptly, a sensation of incomplete bladder emptying and nocturia (Abrams, 1995). More than 30% of the men older than 40 years of age do have symptoms which might indicate BPH. This percentage gets higher regarding age: at the age of 85 about 90% of men have to deal with urinary symptoms which may indicate BPH (Jolleys *et al.*, 1994). Mean age at diagnosis is approximately 60 years. However, the high prevalence of symptoms does not indicate a high prevalence of BPH, because the symptoms associated with BPH can result from other urinary conditions such as prostate cancer, urethral stricture and neurogenic bladder (Norman *et al.*, 1994). On the other hand, only about 50% of all microscopically identifiable BPH lesions ever give rise to a macroscopically enlarged prostate. An enlarged prostate may obstruct the urethra leading to

5. Prices of the tests are based on estimates of real costs: Dfl. 31 for culture; Dfl. 83 for histology; and Dfl. 10 for the urease test (price level 1998).

LUTS, although many of the more bothersome symptoms are caused by resulting bladder hypertrophy. Moreover, an enlarged prostate not always causes LUTS.

Patient data

Although in general people tend to use the word BPH, in this application of the probability modifying model we evaluate the test sequence to diagnose of BOO. We used patient data from a population of consecutive patients form the outpatient urology department of the St.Radboud University Hospital Nijmegen. All patients who were included in the study were referred to the hospital by their general practitioner or by urologists from community hospitals because of LUTS. Patients underwent a diagnostic protocol which besides physical examination included several diagnostic tests: transrectal ultrasound evaluation of the size of the prostate (P, Prosvol); free uroflowmetry to determine the maximum flow rate (Q, Qmax), and post void residual urine volume measured by a abdominal ultrasound (R, Resvol). For the sake of the model we determined the cut-off values at 50 ml, 15 ml/sec, and 50 ml, respectively. As the gold standard for presence or absence of BOO, pressure flow study (PFS) was used (Abrams, 1995). The urethral resistance factor was based on the point of maximum

Table 2.6 Post-test probabilities of test combinations of diagnostic tests for Bladder Outlet Obstruction in relation to the treatment threshold probability of 0.50.

P	Q	R	D^+	D^-	post-test probability	threshold
	+	+ (132)	106	26	$P(D^+ \mid P^+, Q^+, R^+) = 0.80$	>
+	(251)	- (119)	94	25	$P(D^+ \mid P^+, Q^+, R^-) = 0.79$	>
(280)	-	+ (9)	7	2	$P(D^+ \mid P^+, Q^-, R^+) = 0.78$	>
	(29)	- (20)	6	14	$P(D^+ \mid P^+, Q^-, R^-) = 0.30$	<
	+	+ (269)	203	66	$P(D^+ \mid P^-, Q^+, R^+) = 0.75$	>
-	(639)	- (370)	198	172	$P(D^+ \mid P^-, Q^+, R^-) = 0.54$	>
(760)	-	+ (26)	6	20	$P(D^+ \mid P^-, Q^-, R^+) = 0.23$	<
	(121)	- (95)	23	72	$P(D^+ \mid P^-, Q^-, R^-) = 0.24$	<

P: prostate volume; Q: maximum flow rate; R: post void residual volume
>: post-test probability higher than treatment threshold probability
<: post-test probability lower than treatment threshold probability
( ): number of patients between parentheses

flow rate and corresponding detrusor pressure For this standard, a cut-off value of 28 cmH_2O was used The data of a total of 1,040 patients were used in our analyses Rosier *et al* (1996) give an extensive description of the patient population and the diagnostic tests performed

Application of the model

Regarding a treatment decision, it should be noted that for BOO a wide range of treatment options are available ranging from watchful waiting and pharmacotherapy to open prostatectomy Each treatment option with its specific possible benefits and potential risks does influence the final treatment decision Besides this, it is advised that the patient should play a central role in determining the need for a specific treatment (McConnel *et al*, 1994) Therefore the treatment threshold could not be determined empirically, nor estimated, and we analysed the model using several treatment thresholds An overview of the possible combinations of test results and the number of patients related to each combination is presented in Table 2 6 As can be seen from this table the post-tests probabilities of having BOO range from 0 24 to 0 80 This range indicates for which pre-test probabilities of BOO testing is useful regarding the treatment/no-treatment decision Using the lowest and the highest value of this range, the no-treat/test threshold and test/treat threshold are calculated to be, respectively, 0 10 and 0 89 Given the pre-test probability of 0 62 in our population (643/1,040), testing is useful Regarding the final column in this table we used a treatment threshold of 0 50

Table 2 7 Number of tests needed and cost of testing, given all possible test sequences for diagnosis of Bladder Outlet Obstruction using a threshold of 0 50

sequence	first test	second test	third test	number of tests	cost of testing[6](Dfl)
P Q R	1,040 * P	1,040 * Q	29 * R	2,109	400,060
P R Q	1,040 * P	1,040 * R	899 * Q	2,979	424,836
Q R P	1,040 * Q	150 * R	35 * P	1,225	278,178
Q P R	1,040 * Q	150 * P	29 * R	1,219	285,695
R Q P	1,040 * R	1,040 * Q	35 * P	2,115	331,578
R P Q	1,040 * R	1,040 * P	899 * Q	2,979	424,836

P prostate volume, Q maximum flow, R post void residual volume

6 Prices (level 1998) are based on estimates of real costs Dfl 128 50 for transrectal ultrasound evaluation of the size of the prostate, Dfl 254 50 for free uroflowmetry to determine the maximum flow rate, and Dfl 60 for measuring the post void residual volume by a abdominal ultrasound (price)

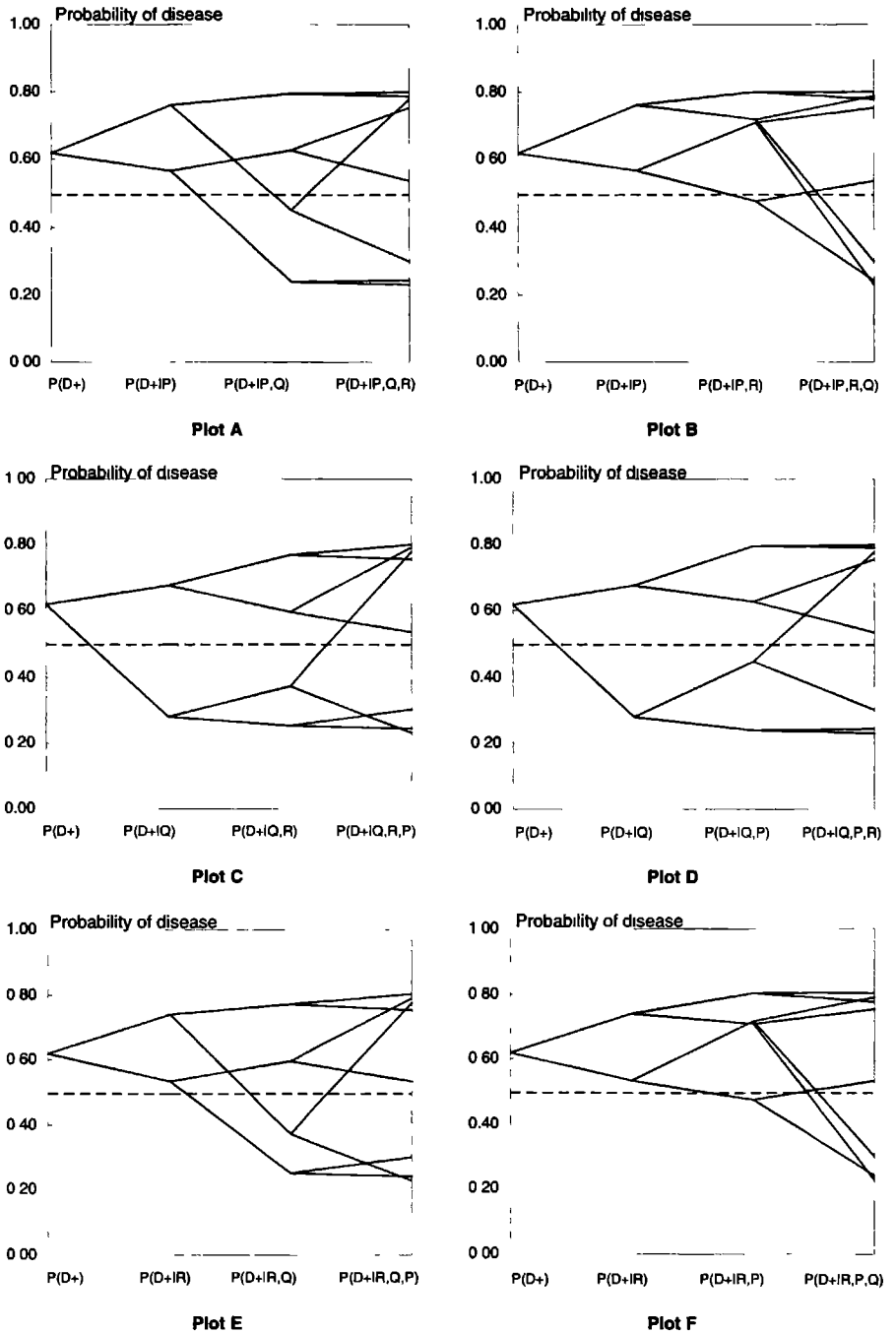


Figure 2.6 Overview of 6-plots showing the different test sequences regarding prostate volume (P), maximum urinary flow rate (Q), and post void residual urine volume (R) for determining the presence of Bladder Outlet Obstruction.

The six sequence strategies regarding diagnostic tests of BOO are shown in Figure 2.6. Again, to decide which test sequence is most efficient, using a predetermined treatment threshold, the number of tests needed and cost of testing is calculated for each sequence. The results when using a threshold of 0.50 are given in Table 2.7. From this table it can be seen that regarding the number of tests needed to finally decide if a patient is indicated to be treated (post-test probability above the treatment threshold) or not to be treated (post-test probability below the treatment threshold) the test sequence Qmax-Prosvol-Resvol is most efficient. Regarding the cost of testing this is the sequence Qmax-Resvol-Prosvol. However, we decided to evaluate the most efficient sequence of testing using a range of treatment thresholds. These results are shown in Table 2.8.

Table 2.8 Most efficient test sequences (number of tests needed and cost of testing) of diagnostic tests for Bladder Outlet Obstruction, regarding different treatment thresholds.

treatment threshold	most efficient test sequence	
	number of tests needed	cost of testing
0.30	Qmax-Prosvol-Resvol	Qmax-Resvol-Prosvol
0.40	Qmax-Prosvol-Resvol	Qmax-Resvol-Prosvol
0.50	Qmax-Prosvol-Resvol	Qmax-Resvol-Prosvol
0.60	Qmax-Prosvol-Resvol	Prosvol-Resvol-Qmax or Resvol-Prosvol-Qmax
0.70	Qmax-Prosvol-Resvol	Prosvol-Resvol-Qmax or Resvol-Prosvol-Qmax
0.80	Prosvol-Resvol-Qmax	Resvol-Prosvol-Qmax

DISCUSSION

In this chapter we described a model to optimise the sequence of diagnostic tests on the basis of efficiency, first, illustrated with data on urinary tract infection. Extensive validation of the test results according to urinary tract infection was beyond the scope of this illustration; no definite conclusions can be drawn about the specific test sequence for this disease. Although the applications of the model regarding diagnosis of Helicobacter Pylori and Benign Prostatic Hyperplasia are based on more appropriate patient data, remarks for discussion have to be made.

Efficiency of the sequence of diagnostic tests was either defined as the least number of tests or the lowest total cost for testing. In fact, the model shows that the different sequences have identical effectiveness defined as the number of persons being misclassified, but the alternative sequences differ regarding the efficiency of the use of diagnostic tests. The probability modifying model is an extension of the existing, but hardly used approach of decision tables (Glasziou & Hilden, 1986). The fact that tests are redundant for subgroups of

patients is clearly shown and of influence on the preferred sequence when efficiency criteria are considered to be important. Other methods to optimise diagnostic test sequences such as the CART method are based on the principle to determine the optimal sequence on the basis of gain in certainty of disease status instead of efficiency (Breiman *et al.*, 1984). Besides this method, a literature review revealed several studies in which sequence of diagnostic tests was studied. Decision analytic models were used regularly to compare alternative sequences of diagnostic tests (Severens & van der Wilt, 1999b), however, it was found that these studies evaluated predetermined sequences without studying all theoretical possible sequences (Kent *et al.*, 1995; Goldberg Kahn *et al.*, 1997; Henschke *et al.*, 1997; Raab & Hornberger, 1997). The latter might have lead to unmanageable bushy decision trees. Examples of evaluation of diagnostic test sequences based on either prospective or retrospective patient data were also found, but again, the test sequence was predetermined without being exhaustively regarding all possible sequences (Chouaid *et al.*, 1993; Einstein *et al.*, 1995). Michel *et al.* (1996) studied the cost-effectiveness of diagnostic strategies in patients with suspected pulmonary embolism using a moddeling approach in which they evaluated all possible combinations of diagnostic tests. However, they assumed independence of the various diagnostic test, which is according to our model, a shortcoming of this approach.

Limitations

The model itself has certain limitations. First, test results are dichotomised and in reality information which is gained by performing a test is often not restricted to a dichotomous answer, thus consequently, information about the status of a patient can be lost. In the case of diagnosis of Helicobacter Pylori, especially the bacterial culture provides more information than only the presence or absence of Helicobacter Pylori. Culture not only answers whether Helicobacter Pylori is present, but also provides information on antimicrobial susceptibility of a patient and this information will allow to treat patients optimally on a individual basis (de Boer, 1997). Considering the other tests involved in this application, histologic examination can not be replaces by a rapid urease test in cases where malignancies must be excluded (Kolts *et al.*, 1993). Besides this, as with most dichotomised tests, the post-test probabilities of disease are dependent on an arbitrary cut-off point (Diamond, 1992). For the BPH application the current evidence regarding uroflowmetry is insufficient to recommend a given cut-off value (McConnel *et al.*, 1994). Besides this, the disease status of a patient has to be dichotomised using a gold standard and the results of the model are highly dependant of the accuracy of this standard. For instance, for the diagnosis of BOO, PFS is considered to be the gold standard. Despite the fact that PFS is an invasive technique (with possible complications), currently it is being considered as a standard diagnostic tool, making it necessary to define a different gold standard for BOO when incorporating PFS in the sequence of tests (Abrams, 1995). A second aspect which can not be incorporated in the model unless a financial translation of waiting time is made, is the turnaround time. In the Helicobacter Pylori

application, a drawback of the bacterial culture is that the turnaround time of diagnosis is rather long (4 to 7 working days) compared to both histology (2 to 3 working days) and rapid urease test (1 to 24 hours) (Kolts *et al.*, 1993). However, regarding Helicobacter Pylori diagnosis, the findings of the probability modifying model are in line with the idea that concordance of two tests leads to the conclusion that, related to any treatment no-treatment threshold, further testing is redundant. The suggestion by Cutler *et al.* (1995) that it may be reasonable to obtain histologic specimens at the time of endoscopy, but not submit them unless the rapid urease test results are negative, is not confirmed by our data and calculations. In contrast to their findings, from our model we have to conclude that there is a diagnostic benefit in performing additional tests after the rapid urease test is positive. Third, as with evaluating medical technologies in general, the selection of the research population is crucial for the result of the probability modifying model. Test characteristics, such as the likelihood ratios, are highly dependent on the patients selected and/or referred for testing and on the overall frequency of abnormal test results in the selected population (Knottnerus & Leffers, 1992b; Schouw *et al.*, 1995; Phelps, 1997). Using data of consecutive patients in several academic and non-academic hospitals might overcome this problem. Fourth, the calculations for optimising the sequence of tests are done for the group of patients as a whole, which leads in principle to the same sequence for all patients, despite the results of a previous test. It might be preferable, once a first test has been performed, to have different sequences for the different subgroups of patients, depending on the first test result. The probability modifying model should be adjusted to overcome this limitation. Fifth, uncertainty about the point estimates of the different post-test probabilities is not incorporated in our model. The uncertainty of the point estimates may not be equal for the points in the plot because these are based on different numbers of patients. In addition, for three tests a sample of 1000 patients is considered to be suitable to use the probability modifying model (Knottnerus, 1992a). However, in our applications it is shown that patients are not equally divided regarding the points in the plot, which, despite the satisfactory number, uncertainty of point estimates is a topic of concern. Sixth, the treatment threshold probability plays an important role in the model. As far as we know, the concept of this threshold has not yet been empirically established. In both the urinary tract example and the applications, the threshold was therefore estimated on the basis of intuitively determined utilities or a wide range of thresholds was used. However, from the different probability modifying plots it can be seen that the post-tests probabilities after three tests are the same, regardless of the sequence of tests. The range between the two post-tests probabilities between which the treatment threshold is defined, can be regarded as the range in which the treatment threshold can be varied without influencing the models' findings analogously to the principles of sensitivity analysis (Glasziou & Hilden, 1986). For the example of the urinary tract infection data, the treatment threshold of 0.67 lies between two post-test probabilities of test result combinations (T_1^+,T_2^+,T_3^-) and (T_1^+,T_2^-,T_3^-), 0.75 and 0.46, respectively. Regarding the Helicobacter Pylori example, the estimated treatment threshold may vary between 0,16 and 0,97 before the conclusions of our model

might change. For the BPH example we used a range of treatment thresholds to show the impact of this threshold on the choice for the most efficient test sequence. Seventh, although it does not seem to be relevant in the application of Helicobacter pylori because all tests need a biopsy specimen, a possible weakness of the model is that the probable (non financial) burden of testing is not taken into account. Risks such as morbidity and mortality related to testing might exist and are not incorporated in the model. Models which incorporate these aspects do exist, but do have their specific difficulties. One method is to translate test morbidity and mortality in monetary terms (Henschke *et al.*, 1997). The other method determines a utility value of the risks of morbidity and mortality, which is reflected in a higher no-treat/test threshold and a lower test/treatment threshold (Sox *et al.*, 1988; Phelps, 1997). The latter method can be incorporated in our model, however, this implicates that for all tests considered a cumulative burden of testing has to be determined.

Despite these limitations, the probability modifying model can be useful for discussing guidelines or protocols concerning diagnostic test strategies and test sequence when efficiency questions are at stake. Further research on the model is needed to handle the current limitations.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. A.J. van Erven, General Hospital Velp, for kindly making available the data regarding the urinary tract infections used to illustrate our model. Dr. W. de Boer, Anna Hospital Oss, is thanked for making available the data on Helicobacter Pylori diagnosis.

CHAPTER 2.2
THE ISSUE OF THE CHOICE OF THE COMPETING ALTERNATIVE: MODELLING OF THERAPEUTIC ALTERNATIVES

Based on Severens JL, Verweij PE, Bos JJ, Donnelly JP & Meis JFGM. Cost-effectiveness of liposomal amphotericin B for the treatment of invasive fungal infections in neutropenic patients: a decision analysis [submitted for publication].

INTRODUCTION

As with the economic evaluation of diagnostic technologies, the question which comparator should be used in the economic evaluation of a therapeutic technology is essential for the study findings. Efficacy trials normally concentrate on a comparator to isolate the actual outcome in a highly controlled context. Placebo controlled trials are therefore widely used. However, in the situation of performing an economic evaluation, effectiveness is to be studied and a placebo is not a relevant decision alternative. Thus, a different comparator should be used in an economic evaluation, such as the technology most effective, cheapest, most frequently used, or any other possible alternative. In case empirical trials are reported in the literature that use comparators that are not highly relevant in a specific decision making context, modelling can be used to investigate alternative comparisons. The main difference between empirical studies and modelling studies is that empirical studies gather information, whereas the modelling studies synthesise information (often-empirical data) without the aim of gathering new empirical data (Brennan & Akehurst, 1999). In this chapter a modelling study is reported that aimed to compare different treatment strategies for patients suspected of having an invasive fungal infection (IFI). The strategy of interest was to our knowledge not reported in the literature.

Invasive fungal infection remains a major cause of morbidity and mortality in the immunocompromised patient (Nemunaitis *et al.*, 1993; Patel *et al.*, 1996; George *et al.*, 1997). Patients who receive cytotoxic treatment for haematological malignancy including those who receive a haemopoietic blood stem cell transplant are especially vulnerable for developing IFI. Candida and Aspergillus species are responsible for the majority of the opportunistic infections and up to 50 % of these patients may develop invasive disease (Denning *et al.*, 1997). The management of IFI is difficult because diagnostic tests are often negative at an early stage of the disease and the choice of antifungal agents is limited. Therefore, it has become standard clinical practice to start antifungal treatment empirical for neutropenic patients with fever who fail to respond despite treatment with broad spectrum antibacterial agents.

In some centres up to 68% of neutropenic patients receive antifungal drugs empirical for persistent fever alone (Goodman *et al.*, 1992; Winston *et al.*, 1993). Usually the treatment of choice is amphotericin B desoxycholate (DC-Amb), while this drug is a fungicidal against a broad spectrum of fungi including Candida and Aspergillus species but is often poorly tolerated. Lipid formulations of amphotericin B including liposomal amphotericin B (AmBisome, Nexstar, St. Odilienberg, The Netherlands), amphotericin B lipid-complex (ABELCET, The Liposome Company, Reeuwijk, The Netherlands) and amphotericin B colloidal dispersion (Amphocil, Zeneca Farma, Ridderkerk, The Netherlands) have been shown to be considerably less toxic than DC-Amb and appear to be equally effective. Yet,

despite their lower toxicity the use of lipid formulations of amphotericin B (L-Amb) is restrained by their high costs. Consequently, L-Amb is only considered for treating those who fail to respond to DC-Amb or those who are intolerant and for patients who develop dose-limiting nephrotoxicity during treatment with DC-Amb (Hiemenz *et al.*, 1998). The results of a recent randomised, double blind trial indicated that empirical treatment of IFI with L-Amb may result in fewer breakthrough fungal infections than occurs with DC-Amb and it was proposed that L-Amb should be considered for first line empirical treatment (Walsh *et al.*, 1999).

Since the costs of L-Amb play such an important role deciding the place of these drugs in treating fungal infections, we undertook a cost-effectiveness analysis. As far as we know the costs and effectiveness of DC-Amb and L-Amb have never been studied prospectively or retrospectively. A method to compare costs and effects of L-Amb and DC-Amb in the face of uncertainty is a decision analytic model. The benefit of a decision analytic model compared to a controlled clinical trial is that for a controlled clinical trial a large, time-consuming and costly study would be required (Drummond *et al.*, 1997). A decision analytic model can be based on data from clinical trials already performed. We compared the cost-effectiveness of using L-Amb for first line empirical treatment of IFI in patients with haematological malignancy with that of DC-Amb by means of decision analytic modelling.

METHODS

Model and Assumptions

Data used in our analysis was limited to studies involving adult patients with haematological malignancies and with a suspicion of IFI. A decision tree was constructed which compared two different treatment strategies. In the first strategy, DC-Amb at a dose of 1 mg/kg/day was given as first line empirical treatment of IFI and was changed to L-Amb for a subset of patients (Figure 2.7, DC/L-Amb strategy). In the other strategy, L-Amb was used for empirical first line therapy of IFI at a dose of 3 mg/kg/day (Figure 2.7, L-Amb strategy). Treatment was considered to be successful when fever subsided (body temperature < 38 °C) and resolution of symptoms resolved whether or not side-effects occurred.

For each strategy a limited number of possible side effects of treatment with DC-Amb or L-Amb was included in the decision tree. Only side effects which are common or which have a probability of mortality were used in the model. Furthermore, the need for co-medication or extra hospitalisation for the side effects was incorporated in the decision tree. The side-effects which met these criteria included fever (body temperature of 38°C or more) and/or chills related to the infusion of the drug (Walsh *et al.*, 1997; Prentice *et al.*, 1997; Walsh *et al.*,1999; Ellis *et al.*, 1998), nephrotoxicity (as a rise in the serum creatinine level of

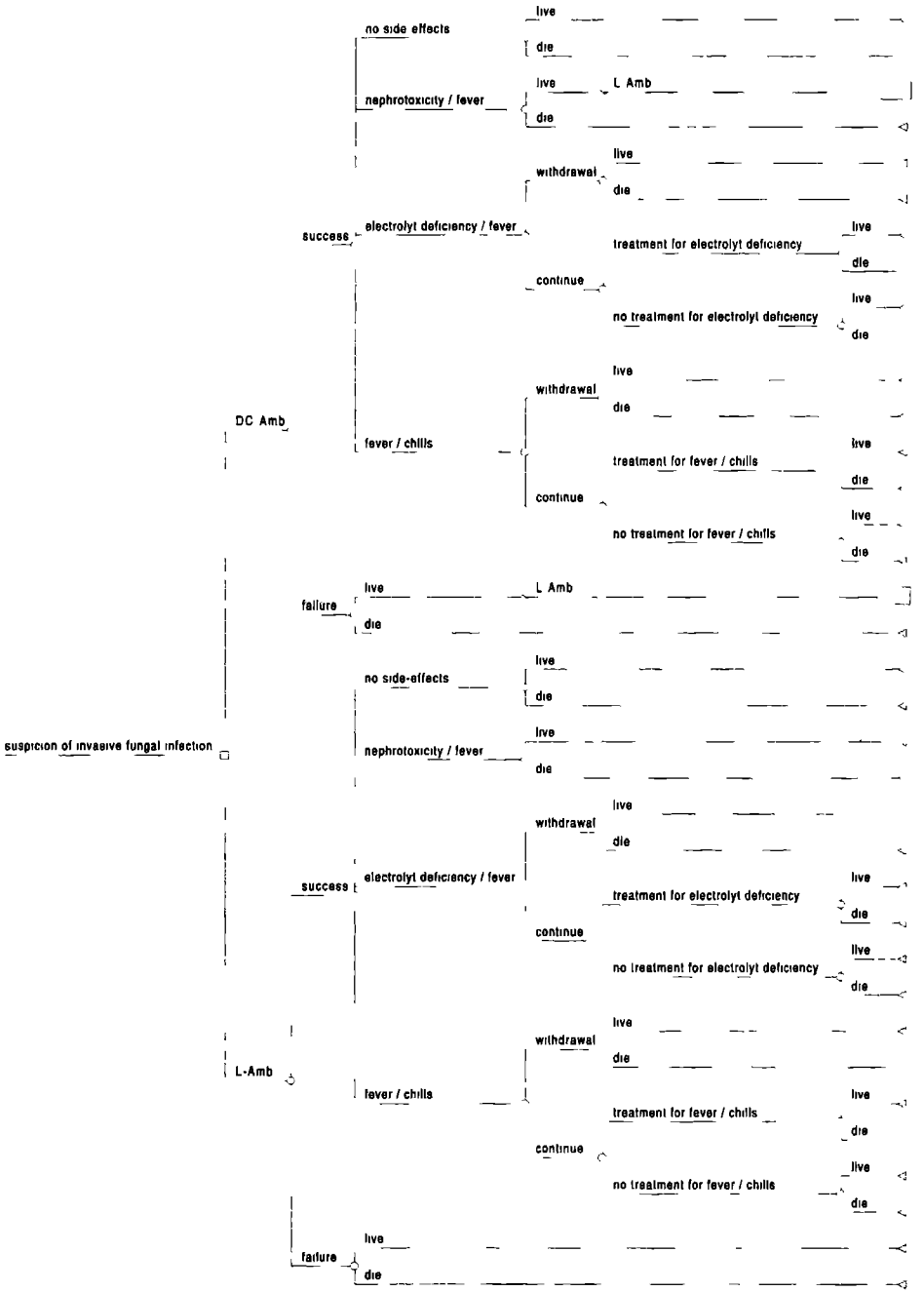


Figure 2.7 Decision tree for two strategies for the treatment of suspicious invasive fungal infection (IFI). In the first strategy, amphotericin B desoxycholate (DC-Amb) was given as first line empirical treatment of IFI and was changed to liposomal amphotericin B (L-Amb) in case of a treatment failure or nephrotoxicity (DC/L-Amb strategy). In the other strategy, L-Amb was used for empirical first line therapy of IFI (L-Amb strategy).

at least twice the upper limit (Walsh *et al.*, 1997; Prentice *et al.*, 1997; Walsh *et al.*, 1999) of normality), and electrolyte deficiency. In the DC/L-Amb strategy, first line treatment of suspected IFI was DC-Amb. DC-Amb was discontinued and changed to L-Amb, when treatment failed whereas L-Amb was continued in the L-Amb strategy. This was included in the model. When nephrotoxicity was assumed to lead to extra hospitalisation in both strategies and change to L-Amb was modelled in the decision tree. The occurrence of either fever or chills or electrolyte deficiency could result in either continuation of the drug in both strategies and the administration of co-medication, or stopping the drug altogether.

The probability of death due to the underlying disease was considered to be equal for both strategies. An additional cause of mortality was death due to a proven IFI in case of a treatment failure. Treatment failure was defined as progression of IFI during treatment with either DC-Amb or L-Amb, and the probability of treatment failure was presumed to differ between both strategies. Another additional cause of mortality included in the model was related to side effects due to treatment with DC-Amb or L-Amb.

Data and probabilities

MEDLINE was searched to retrieve relevant publications, using the search criteria amphotericin B desoxycholate, liposomal amphotericin B, invasive fungal infection, and haematologic. In addition, articles were located from the bibliographies of the papers retrieved and published abstracts of randomised clinical trials were included. Articles and abstracts were then selected that reported the following criteria: 1) treatment with DC-Amb or L-Amb of adult patients with haematological malignancies; 2) the number of cases in which IFI was suspected before giving treatment; 3) and the number of patients who died. All articles that reported the results of studies performed in mixed patients groups without making a distinction between those infected with HIV+ or solid organ transplant recipients and those with haematological malignancies were excluded.

Data were extracted by first calculating the number of successfully treated cases (total number of cases minus the treatment failures) then the number of cases without side effects was determined. If the number of patients without side effects was not explicitly mentioned in the article, the number was calculated by subtracting the number of cases of nephrotoxicity, electrolyte deficiency and fever or chills from the total number successfully treated patients. Although most papers provided data on the occurrence of individual side effects, the number of cases that experienced more than one side effect was generally not reported. In order to prevent overestimating of the total number of patients affected by side effects, we made the assumption that patients who experienced nephrotoxicity or electrolyte deficiency also experienced fever or chills. Cases of fever or chills that could not be accounted for as nephrotoxicity or electrolyte deficiency were allocated to fever or chills alone. We considered fever or chills to have no impact on effectiveness and hardly any impact on costs. The data were pooled once the number of patients was determined for the different categories for each

study, by calculating the probabilities of successful treatment, treatment failure, and occurrence of side-effects, based on the total number of cases of the joint studies. Since the probability of mortality due to underlying disease was considered to be similar for both treatment strategies, these probabilities were pooled over all data. The probability of treatment failure and the probability of mortality due to failure of treatment for proven IFI differed between both strategies and was calculated exclusively with the results of data from randomised trials (Prentice *et al.*, 1997; Walsh *et al.*, 1999). Finally, it was assumed that events not reported in the article or abstract did not actually occur.

Costs

The costs were assumed to have been incurred from the time of starting treatment until discharge from the hospital and included the costs of antifungal treatment and those associated with the occurrence of side-effects. Costs were calculated in Dutch guilders and exchanged to US dollars (USD) using the mean exchange rate of 1997. The total costs of treatment with DC-Amb and L-Amb were based on the cost per day and the duration of treatment. The duration of treatment with both DC-Amb and L-Amb was set at 28 days on the basis of expert opinion and chart review of patients treated for IFI at the Department of Haematology of the University Hospital Nijmegen. The dose of DC-Amb was set at 1 mg/kg/day for a patient weighing 70 kg and the cost of the drug was estimated to be USD 23.2 per day based on retail prices in the Netherlands which averaged USD 33.1 per 100 mg. The cost of L-Amb per day was based on two drugs available in the Netherlands; Ambisome and Amphocil. Both are sold for the same price of USD 231 per 50 mg which, for a 70-kg patient given a dose of 3 mg/kg/day amounted to a daily cost of USD 969.

Nephrotoxicity due to treatment with DC-Amb or L-Amb was assumed to increase the length of stay (LOS) by seven extra days (Hiemenz *et al.*, 1998), which amounted to USD 4,648. A single day of hospitalisation in an academic hospital was estimated at USD 664 (Rutten *et al.*, 1993). In addition, it was assumed that due to electrolyte deficiency the LOS increased by seven days. Infusion related reactions such as fever or chills were assumed to take place throughout treatment and pethidine at a dose of 25 mg was assumed to be given as treatment costing USD 0.6/day.

Analysis

The expected costs and expected effectiveness expressed as lives saved were calculated for each treatment strategy using a differential approach, thus concentrating on th costs that were expected to differ between the strategies. The sum of either the cost or the effect for each of the branches was multiplied by their probability of occurrence.

A cost-effectiveness analysis was performed to compare the two strategies. Given the nature of the clinical problem, the time frame considered in this analysis was limited to the

LOS. The cost-effectiveness analysis was performed from the perspective of the health-care system, which implied that only direct medical costs were incurred. The strategy yielding the greater effectiveness at a lower cost was considered to be dominant. If this was not the case and one strategy was both more effective and more costly, the relationship between costs and effectiveness was expressed in an incremental cost-effectiveness ratio. Incremental cost-effectiveness ratios were calculated by dividing the additional costs of the L-Amb strategy compared to DC/L-Amb by the effectiveness gained, and was expressed in terms of additional US dollars per life gained.

First, the cost-effectiveness analysis was performed using baseline values for probability and cost estimates. Next we varied these estimates in the decision tree by using a sensitivity analysis. This enabled us to explore for which probabilities or costs the results of the analyses were robust and for which they were sensitive regarding the range used in the sensitivity analysis. The ranges of probabilities were determined by calculating 95%

Table 2.9 Summary of clinical studies of amphotericin B desoxycholate (DC-Amb) for the treatment of invasive fungal infections (IFI).

reference	treatment of underlying disease	number of cases	doses	treatment failures	death due to proven IFI in case of treatment failure	mortality
Walsh 1999	haematological malignancies, solid tumors	344	0.6 mg/kg/day	43	11	25
Prentice 1997	leukemia,lymphoma	39[#]	1 mg/kg/day	21[§]	0	9
Anaissie 1996	BMT,leukemia,cancer	67	25-50 mg/day	20	4	5
Pascual 1995	BMT, blood stem cell	10	50 mg/day	1	0	0
Ellis 1995	leukemia,Hodgkin's,BMT	25	0.5 mg/kg/day	3	2	3
Richard 1993	BMT	8	0.5 mg/kg/day	0	0	4
Tollemar 1992	haematological malignancies/ lymphoma	9[&]				7
Blade 1992	leukemia	10	mean 4.6 g / 18.3 days	4	0	4
total		512		92	17	57

\# data is based on adult patients
§ not responding to the drug
& organ transplant patients excluded

confidence intervals. The ranges of the costs used in the sensitivity analysis were calculated according the following principles. The baseline value of cost of DC-Amb was estimated at USD 649.9. The range of these costs was estimated on the difference in duration of use 1-35 days (USD 23.2-811.9). Similarly, the baseline value of cost of L-Amb was estimated at USD 21,137. The sensitivity analysis was done in two ways. The first sensitivity analysis was based on a difference in prices of L-Amb (USD 10 to 1,000), for a fixed duration of 28 days amounting to (USD 280 to 28,000). The second sensitivity analysis was based on the difference in duration of use 1-35 days based on a fixed cost of USD 969 resulting in (USD 969 to 33,915). The baseline value of costs of treatment for both nephrotoxicity and electrolyte deficiency were estimated at USD 4,649 for the increased LOS and sensitivity analysis was based on 0-14 days extra hospitalisation yielding an extra cost of USD 0 to 9,298.

The model was generated using DATA version 2.6 software (TreeAge Software Inc., Boston USA) for Windows (Microsoft Corporation).

RESULTS

Literature review

Fourteen articles met our criteria of which four reports provided data from treatment with DC-Amb in open noncomparative trials (Blade *et al.*, 1992; Richard *et al.*, 1993; Ellis *et al.*, 1995; Anaissie *et al.*, 1996;), six dealt with treatment with L-Amb (Tollemar *et al.*, 1990; Chopra *et al.*, 1991; Chopra *et al.*, 1992; Krüger *et al.*, 1995; Oravcová *et al.*, 1995; Ellis *et al.*, 1998) of which only one was a randomised study comparing 1 mg/kg/day with 4 mg/kg/day L-Amb (Ellis *et al.*, 1998). The remaining four publications reported a comparison of treatment with DC-Amb with L-Amb of which two were randomised clinical trials (Prentice *et al.*, 1997; Walsh *et al.*, 1999), the other two being non-randomised comparative studies (Tollemar *et al.*, 1992; Pascual *et al.*, 1995). Two studies included patients with haematological malignancies and reported solid tumours (Oravcová *et al.*, 1995; Ellis *et al.*, 1998) and two studies reported on patients with lymphomas (Ellis *et al.*, 1995; Prentice *et al.*, 1997). The fourteen studies reported the results of treating 512 patients with haematological malignancies receiving treatment with DC-Amb (Table 2.9) and 579 similar patients receiving treatment with L-Amb (Table 2.10). An overview of the reported side effects and the treatment of these side effects of DC-Amb and L-Amb are shown in Tables 2.11 and 2.12 respectively. Based on these figures, pooled data were calculated and are shown in Table 2.13. These were used for baseline estimates of the probabilities used in the decision tree.

Table 2 10 Summary of clinical studies of lipid formulations of amphotericin B (L-Amb) for the treatment of invasive fungal infections (IFI)

reference	treatment of underlying disease	number of cases	doses	treatment failures	death due to proven IFI in case of treatment failure	mortality
Walsh 1999	haematological malignancies, solid tumors	343	3 mg/kg/day	43	4	21
Ellis 1998	BMT, leukemia, solid tumor	46	4 mg/kg/day			25
Prentice 1997	leukemia, lymphoma	47[#]	3 mg/kg/day	17[§]	0	3
Oravcova 1995	haematological malignancies, solid tumors	20	5 mg/kg/day			3
Pascual 1995	haematological malignancies	10	50 mg/day			0
Kruger 1995	BMT, blood stem cell	50	2 8 mg/kg/day			16
Chopra 1992	BMT, chemotherapy	31	5 mg/kg/day			14
Tollemar 1992	haematological malignancies/ lymphoma	8[&]				3
Chopra 1991	haematological malignancies	20	3 mg/kg/day			10
Tollemar 1990	BMT	4				2
total		579		60	4	97

\# data is based on adult patients
§ not responding to the drug
& organ transplant patients excluded

Table 2 11 Summary of reported side-effects and treatment of these side-effects of amphotericin B desoxycholate (DC-Amb)

	Walsh 1999	Prentice 1997	Anaissie 1996	Pascual 1995	Ellis 1995	Richard 1993	Tollemar 1992	Blade 1992	total
successfully treated	301	18[#]	47	9	22	8	9[§]	6	420
no side-effects	150	10	20	1	3	5	0	3	192
nephrotoxicity / possibly fever	117	6	22		1	3	8	0	157
electrolyte deficiency / possibly fever		2	5	8				3	18
withdrawn due to electrolyte deficiency			0	0				1	1
treatment for electrolyte deficiency			0	8				0	8
fever and/or chills	34		0	0	19		1		54
withdrawn due to fever and/or chills			0	0	0		0		0
treatment for fever and/or chills			0	0	19		0		19

data is based on adult patients
§ organ transplant patients excluded

Table 2.12 Summary of reported side-effects and treatment of these side-effects of lipid formulations of amphotericin B (L-Amb).

	Walsh 1999	Ellis 1998	Prentice 1997	Oravcova 1995	Pascual 1995	Kruger 1995	Chopra 1992	Tollemar 1992	Chopra 1991	Tollemar 1990	total
successfully treated	300	46	30[#]	20	10	50	31	8[§]	20	4	519
no side-effects	180	41	27	15	3	48	27	8	20	4	373
nephrotoxicity / possibly fever	65	5	3					0			73
electrolyte deficiency / possibly fever			0	5	7		4				16
withdrawn due to electrolyte deficiency				0	0		0				0
treatment for electrolyte deficiency					7		4				11
fever and/or chills	55			0	0	2		0			57
withdrawn due to fever and/or chills				0	0	1					1
treatment for fever and/or chills				0	0	1					1

\# data is based on adult patients
§ organ transplant patients excluded

Cost-effectiveness

The cost estimates used in our model are listed in Table 2.14. Analysing the decision tree using the baseline values, first-line treatment with L-Amb proved to be the most effective resulting in an expected probability for survival of 0.853 compared with 0.771 for the DC/L-Amb strategy. However the L-Amb strategy was more effective, it was also more expensive than the DC/L-Amb strategy, thus no strategy was dominant to the other. The expected cost per patient of L-Amb was USD 27,810 while the expected cost per patient of DC/L-Amb was USD 12,776. Note that the last cost estimate includes not only the costs of DC-Amb but also the costs of L-Amb used in case of a treatment failure or when there was nephrotoxicity. To calculate the incremental cost-effectiveness ratio of the L-Amb strategy compared to the DC/L-Amb strategy, the additional cost (USD 27,810 – USD 12,776 = USD 15,034) of the L-Amb strategy was divided by the probability of saving a life (0.853-0.771 = 0.082). This calculation yielded a ratio of nearly USD 183,000 per life saved.

Table 2.13 Probabilities of side-effects of the treatment, the treatment of these side-effects, and mortality used in the model.

	DC-Amb		L-Amb	
	pooled probabilities	calculated range[#]	pooled probabilities	calculated range[#]
morbidity				
treatment failure	0.18	0.15-0.21	0.10	0.08-0.12
no side-effects	0.46	0.41-0.51	0.72	0.68-0.76
nephrotoxicity / fever	0.37	0.32-0.42	0.14	0.11-0.17
electrolyte deficiency / fever	0.04	0.02-0.06	0.03	0.02-0.05
treatment withdrawn due to electrolyte deficiency	0.06	0-0.17	0	0
treatment for electrolyte deficiency	0.44	0.21-0.67	0.69	0.51-0.87
fever and/or chills	0.13	[§]	0.11	[§]
treatment withdrawn due to fever and/or chills	0	0	0.02	0-0.06
treatment for fever and/or chills	0.35	0.22-0.48	0.02	0-0.06
mortality				
mortality due to underlying disease[&]	0.14	0.12-0.16	0.14	0.12-0.16
additional mortality due to proven invasive fungal infection in case of treatment failure	0.18	0.1-0.26	0.07	0.05-0.09
additional mortality due to nephrotoxicity[‡]	0.001	0-0.01	0.001	0-0.01
additional mortality due to electrolyte def.[‡]	0.001	0-0.01	0.001	0-0.01

\# $CI \rightarrow 95\% = p \pm 1.96\sqrt{\frac{p(1-p)}{n}}$

§ fever is the node in the model with a probability estimate dependent on the other probabilities and therefore it can not be used in a sensitivity analysis

& pooled over all data

‡ set by expert opinion

Sensitivity analysis

Sensitivity analyses performed over all calculated ranges of the variables in the model showed that two variables had an important impact on the findings of our decision model namely, the cost of L-Amb per day and the duration of treatment with the drug. Changing the cost of L-Amb per day showed that if L-Amb cost less than USD 121.6 a day the L-Amb strategy was more effective than the DC/L-Amb strategy in terms of lives saved at a lower cost. If L-Amb cost more than USD 121.6, L-Amb was more effective at a higher cost. Varying the duration of treatment with L-Amb showed that when treatment exceeded 4 days it was more effective but at a higher cost. L-Amb always yielded a higher effectiveness at a higher cost than DC/L-Amb for all the other cost variables explored in the sensitivity analyses, with a maximum incremental cost-effectiveness ratio of USD 235,000 per life saved (Table

Table 2 14 Baseline cost estimates in US dollars (USD) and ranges for the sensitivity analyses used in the model

	baseline value (USD)[#]	range (USD)
amphotericin B desoxycholate (1 mg/kg/day)	649 (28 days)	23-812 (1-35 days)
liposomal amphotericin B (3 mg/kg/day)[§]	27 137 (28 days)	969-33,921 (1-35 days)
extra hospitalisation due to nephrotoxicity or electrolyte deficiency	4,649 (7 days)	0-9,298 (0-14 days)
fever and/or chills treatment (pethidine 25 mg/day)	17 (28 days)	13-22 (21-35 days)

the mean exchange rate for 1997 from US dollars to Dutch guilders was 1 95
§ based on the average prices of Amphocil and AmBisome

2 15) Sensitivity analyses over the probabilities of side effects had no significant impact on the incremental cost-effectiveness ratios Assuming a probability of failure of treatment with DC-Amb of 0 15 demonstrated an incremental cost-effectiveness ratio of USD 204,000 per life saved and USD 166,000 per life saved if the probability was set at 0 21 Mortality due to underlying disease demonstrated an incremental cost-effectiveness ratio of USD 196,000 per life saved if the probability was calculated at 0 12 and USD 174,000 per life saved when it was calculated at 0 16 Additional mortality due to a proven IFI in case of treatment failure of DC-Amb demonstrated an incremental cost-effectiveness ratio of USD 210,000 per life saved when the probability was 0 1 and USD 164,000 per life saved when the calculated probability was 0 26 By contrast, the additional mortality due to a proven IFI that failed treatment with L-Amb had barely any impact with an incremental cost-effectiveness ratio of USD 181,000 and USD 186,000 per life saved when the calculated probability was 0 05 and 0 09 respectively

Discussion

The costs of liposomal formulations of amphotericin B play a decisive role in determining their place in the treatment of IFI in neutropenic patients Indeed the high costs of treatment with these drugs for empirical antifungal therapy has led in many institutes to reserve them for those patients who either fail to respond to the conventional formulation of the drug or to who develop dose-limiting nephrotoxicity In order to explore the feasibility of using L-Amb for first line empirical treatment of IFI we compared the effectiveness and expected costs of such strategy with empirical therapy with DC-Amb using decision analysis Within all the ranges explored, L-Amb remained the most effective strategy increasing the probability of survival

Table 2 15 Results of the one way sensitivity analyses incremental cost-effectiveness ratios (ICER) of using lipid formulations of amphotericin B (L-Amb) versus amphotericin B desoxycholate (DC-Amb) as first line empirical treatment of for patients suspected of invasive fungal infections (IFI), in 1,000 US dollar (USD) per life saved

	range of values	ICER (USD 1,000/life saved)
baseline values[#]		183
probability of treatment failure		
DC-Amb	0 15 - 0 21	204 – 166
L-Amb	0 08 - 0 12	182 – 185
probability of nephrotoxicity		
DC-Amb	0 32 - 0 42	211 – 159
L–Amb	0 12 - 0 18	183 – 185
probability of electrolyte deficiency		
DC-Amb	0 02 - 0 06	184 – 183
L-Amb	0 02 - 0 05	183 – 184
probability of treatment withdrawn due to electrolyte deficiency		
DC-Amb	0 - 0 17	183 – 184
probability of treatment for electrolyte deficiency		
DC-Amb	0 21 - 0 67	184 – 183
L-Amb	0 51 - 0 87	183 – 184
probability of treatment withdrawn due to fever and/or chills		
L-Amb	0 - 0 06	183 – 183
probability of treatment for fever and/or chills		
DC-Amb	0 22 - 0 48	183 - 183
L-Amb	0 - 0 06	183 - 183
probability of no side-effects		
DC-Amb	0 41 - 0 51	166 - 202
L-Amb	0 68 - 0 76	184 - 183
probability of mortality due to underlying disease	0 12 - 0 16	196 - 174
probability of additional mortality due to proven IFI in case of treatment failure		
DC-Amb	0 1 - 0 26	210 - 164
L-Amb	0 05 - 0 09	181 - 186
probability of additional mortality due to nephrotoxicity		
DC-Amb	0 - 0 01	184 - 179
L-Amb	0 - 0 01	183 - 185

Table 2 15 continued

probability of additional mortality due to		
electrolyte deficiency		
DC-Amb	0 - 0 01	183 - 185
L-Amb	0 - 0 01	183 - 184
costs[&]		
DC-Amb	USD 23 - USD 812	191 - 181
L-Amb in time use (days)	1 - 35 (days)	DOMINANT[§] - 235
L-Amb in costs	USD 10 - USD 1000	DOMINANT - 190
add hospitalisation due to nephrotoxicity	USD 0 - USD 9,298	196 - 171
add hospitalisation due to electrolyte deficiency	USD 0 - USD 9,298	184 - 183
treatment for fever or chills/chills	USD 13 - USD 22	183 - 183

\# see Table 2 13 for more details on the values used in the baseline analysis
§ the L-Amb strategy is dominant compared to the DC/L-Amb strategy, thus, the L-Amb strategy yielded a higher effectiveness at lower costs
& see also Table 2 14 for more details on the values used in the baseline analysis and the sensitivity analyses

by 8% but it was also a more expensive strategy than DC/L-Amb costing at least USD 15,000 more per treated patient However, the additional cost per life saved exceeded USD 235,000 over all calculated ranges of the variable values explored which corresponds to USD 23,500 per life-year saved for a patient with a remaining life expectancy of 10 years The costs and the duration of treatment with L-Amb was dominant provided L-Amb cost less than USD 121 6 per day in which case the strategy was more effective at a lower price Thus, reducing the price of L-Amb can only have positive impact on implementing a strategy, which involves giving L-Amb as first line treatment rather than initially starting with DC-Amb The sensitivity analyses also showed that varying probabilities of side effects had only a negligible impact on the cost-effectiveness ratios

Furthermore, in a two-way sensitivity analysis the complementary impact of varying both cost per day and the duration of treatment of L-Amb was studied We conclude that the costs per day or the duration of use of L-Amb has to decrease considerably before L-Amb becomes the dominant strategy If L-Amb cost USD 200 (baseline was USD 969) and the duration of use was 18 days (baseline was 28 days) L-Amb was the preferred strategy (Figure 2 8)

Ellis *et al* (1998) performed a randomised controlled trial comparing treatment with 1 mg/kg/day L-Amb and 4 mg/kg/day L-Amb and found no significant difference between the two groups in the number of side effects We therefore undertook another analysis assuming that L-Amb was given at a 1-mg/kg/day dose instead of 3 mg/kg/day thereby reducing the daily costs of the product This yielded an incremental cost-effectiveness ratio of nearly USD 68,000 per life saved Which corresponds to USD 6,800 per life year saved for a patient with a remaining life expectancy of 10 years This is an amount that is considered acceptable

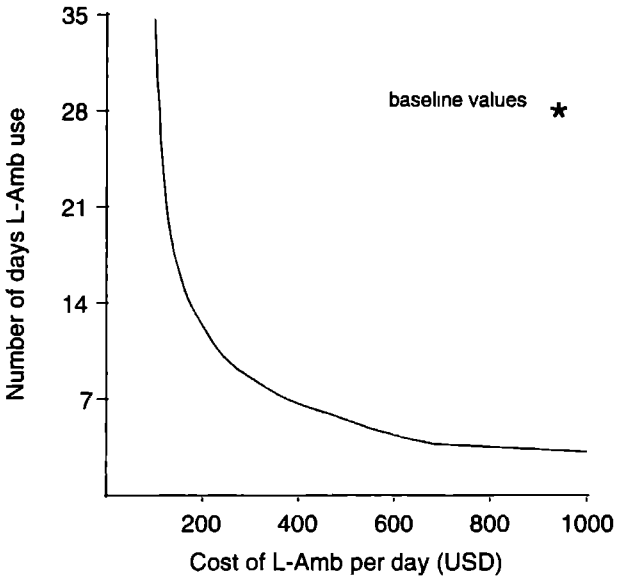


Figure 2.8 Two-way sensitivity analysis of the decision tree to determine the effect simultaneously varying of costs per day and duration of use of L-Amb In this figure it is clear that L-Amb became the preferred strategy if the product became cheaper and the duration of use became shorter. The baseline values are represented with a *. Every point under the curve indicates a higher effectiveness in lives saved and lower costs (dominance), every point above the curve indicates a higher effectiveness in lives saved and higher costs for the L-Amb strategy compared to the DC/L-Amb strategy.

(Goldman *et al.*, 1992). Although the L-Amb strategy did not reach dominance it induced a considerable reduction in incremental costs per life saved compared to the analysis done over 3 mg/kg/day, which was USD 183,000 per life saved.

Theoretically, the probability of mortality due to the underlying disease should be the same for both strategies and therefore we pooled the estimated probability over all data for our analysis resulting in a probability of 0.14. However, after pooling the data of DC-Amb and L-Amb separately we found a probability of mortality due to the underlying disease of 0.11 for the DC-Amb group and of 0.17 for the L-Amb group (Tables 2.9 and 2.10). In the studies which evaluated the efficacy of L-Amb a number of patients were treated first with DC-Amb before switching to L-Amb but, it was not possible to separate these patients from those who received first line therapy with L-Amb. It is therefore possible that the patients who were first treated with DC-Amb differed from those who received first line treatment with L-Amb with respect to their probability of mortality due to the underlying disease. Thus, in case the

probability of mortality due to the underlying disease had been pooled in the same manner as the other probability estimates, this might have introduced selection bias.

The probability of treatment failure and of mortality due to a proven IFI in case of a treatment failure also suffers from possible selection bias. It is still possible that these probabilities are higher for the patients who initially received DC-Amb than for those who received first line treatment with L-Amb. Therefore these probabilities were calculated on the basis of data derived exclusively from randomised clinical trials (Prentice *et al.*, 1997; Walsh *et al.*, 1997).

None of the studies we retrieved made a distinction between patients who had only one side effect and those who had more than one side-effect which was apparent from the fact that summation of all probabilities of side-effects of L-Amb or DC-Amb reported, resulted in a probability greater than 1.

The study by Prentice *et al.* (1997) enumerated the drug-related adverse events which allowed us to estimate accurately the number of patients who had no side-effects. In addition, it was also possible to estimate accurately those who developed nephrotoxicity and electrolyte deficiency. However, simple addition of the number of patients exceeded those successfully treated minus the number of those without side effects so there was double counting of nephrotoxicity and electrolyte deficiency. It is possible that a side effect occurred at the same time as treatment was deemed to have failed since failure was defined as no response to the drug. Therefore, since it was not clear in any of the selected studies, this information was not included in our decision tree. In the article by Prentice *et al.* (1997) at least one side effect could have occurred while treatment failed. This is possibly also the case in the other retrieved studies. However, this omission is likely to have limited consequences for our analyses because all side effects with a significant impact were included as treatment success.

Possible dose-response relationships were not taken into consideration in our decision analytic model; the results might well have been different if these had been taken into account. A mean of 3.3 mg/kg/day for L-Amb was estimated from the selected studies. Which is a slightly higher than the assumed 3 mg/kg/day and the dose of DC-Amb was 0.6 mg/kg/day which is less than the 1 mg/kg/day we employed. However, neither the efficacy nor the toxicity of treatment with amphotericin B appears dose dependent (Ellis *et al.*, 1998), so these deviations may not have had any consequences.

The probability of mortality due to a proven IFI in case of a treatment failure is rather low in the retrieved studies. Presumably, patients who were considered to have a fungal infection while treated with L-Amb or DC-Amb were not always infected because the probability of mortality due to a proven IFI in case of a treatment failure would be much higher than the probability found; 0.18 for DC-Amb and 0.15 for L-Amb.

In conclusion, our decision analytic model yielded a higher effectiveness in lives saved if L-Amb is used as first line treatment compared to initial treatment with DC-Amb but that, as expected, the L-Amb strategy was also more expensive. Since this is highly dependent upon the cost and the duration of treatment with liposomal amphotericin B, L-Amb could only

become the drug of first choice for the empirical treatment of neutropenic patients if the cost of this product is decreased dramatically.

CHAPTER 3
THE ISSUE OF THE RELEVANT COSTS AND CONSEQUENCES

CHAPTER 3.1
THE ISSUE OF THE RELEVANT COSTS AND CONSEQUENCES: DETERMINING THE TIME HORIZON OF THE ANALYSIS

Based on Severens JL, Donnelly JP, Meis JFGM, Vries Robbé PF de, Pauw BE de & Verweij PE (1997). Two strategies for managing invasive aspergillosis: a decision analysis. *Clinical Infectious Diseases* 25: 1148-1154.

INTRODUCTION

The process of diagnosis focuses on the reduction of uncertainty about the presence of a disease of a person (Elstein *et al.*, 1993). Hence, the number of cases detected is an effectiveness parameter which can be used in economic evaluations. Because, normally, the results of an economic evaluation are expressed in a cost-effectiveness ratio, this leads to the costs per case detected. However, this parameter can be regarded as an intermediate outcome measure in the sense that it does not reflect the actual health outcome as a result of diagnosis and eventual treatment of a patient. In the literature review it was shown that this intermediate outcome is used regularly (18.3% of the studies in which a ratio was defined) (Severens & van der Wilt, 1999b). Besides this, cost per 'specific measure' was found in 31.4% of the studies mentioning a ratio. The latter category was often used in studies which evaluated screening programmes. Ratios like 'costs per case of infection prevented', 'costs per born baby with abnormalities prevented', but also 'cost per treated patient' were used. Thus, this category can also be regarded as studies using intermediate outcomes. In fact, 49.7% of the studies mentioning a ratio used intermediate effectiveness measures which does not reflect, as mentioned above, the actual health outcome as the main objective of medical interventions.

Determining the effectiveness measure used in an economic evaluation is related to the relative time span in which effectiveness is measured, in other words, the time horizon of analysis. The time horizon used when evaluating efficiency of an intervention in general should extend far enough in the future to capture the major health and economic outcomes (Torrance *et al.*, 1996). This chapter aims to stress the importance of determining the time horizon of a study when evaluating diagnostic tests.

A second topic of this chapter concerns modelling. Modelling techniques can be useful, in case it is not possible to prospectively measure health outcome when evaluating a diagnostic technology because of the duration of the study and an intermediate outcome measure has to be used (Canadian Coordinating Office for Health Technology Assessment, 1997). The mentioned aspects are illustrated with diagnosis and treatment of invasive aspergillosis.

Invasive aspergillosis

Invasive pulmonary aspergillosis is a life-threatening complication of intensive chemotherapy for malignancies, chronic treatment with high dose corticosteroids, transplantation, and AIDS. Treating this infection as early as possible is one of the prerequisites for a favourable outcome (Aisner *et al.*, 1977), but since no reliable means of early diagnosis exists, it has become commonplace to treat patients at high risk empirically with amphotericin B when there is the slightest clinical suspicion of invasive aspergillosis (Denning, 1994). As many as 68% of

neutropenic patients will receive amphotericin B empirically, often for persistent fever alone (Goodman *et al.*, 1992; Winston *et al.*, 1993), even though the desoxycholate formulation of amphotericin B (DC-Amb; Fungizone, Bristol Myers Squibb, Woerden, the Netherlands) is associated with adverse drug reactions such as fever, rigors, and impaired renal function. In addition, the incidence of disease for most groups at risk is ~10%. Therefore, alternative strategies are needed for the early recognition of invasive aspergillosis to allow better selection of patients who need treatment, while reducing the number of patients who are exposed unnecessarily to these drugs and their side effects.

In addition, better selection of patients who require treatment for invasive aspergillosis will also help to contain the cost of treatment when novel and expensive drugs are used. For instance, lipid-formulations of amphotericin B, such as liposomal amphotericin B (L-Amb; Ambisome, Nexstar, St. Odilienberg, the Netherlands), amphotericin B lipid-complex (Abelcet, The Liposome Company, Reeuwijk, the Netherlands), and amphotericin B colloid dispersion (Amphocil, Zeneca Farma, Ridderkerk, the Netherlands), are available and are being used increasingly for treating invasive aspergillosis because they induce fewer side effects than occur with the same dose of DC-Amb. Although equal or improved efficacy of lipid formulations of amphotericin B in treating invasive aspergillosis has not been proven, all the studies to date have shown that lipid-formulations of amphotericin B are at least as effective as DC-Amb, even in those patients who did not respond (Hiemenz & Walsh, 1996; Ng TTC & Denning, 1995; Hiemenz *et al.*, 1995). Moreover, clinical trials are ongoing to establish the efficacy of lipid-formulations of amphotericin B as empirical therapy for fever and neutropenia and as first-line therapy for documented aspergillosis (Hiemenz & Walsh, 1996).

However, the cost of treatment with these formulations is high. L-Amb is as much as 40 times more expensive than DC-Amb and is only considered for managing invasive aspergillosis in patients with impaired renal function or in those who are intolerant of DC-Amb or for managing progressive disease despite treatment with the desoxycholate formulation of the polyene (Hiemenz & Walsh, 1996). Even if desirable, empirical treatment with L-Amb is simply not feasible since resources for health care are finite and there is increasing pressure to contain costs despite a indisputable rise in demand.

To address this issue we devised a diagnostic approach (alternative strategy) that incorporates *Aspergillus* antigen detection, high resolution CT scanning, and radionuclide imaging with-indium-111-labelled human IgG (^{111}In-IgG). Each of these procedures have been shown to contribute to establishing the diagnosis of invasive aspergillosis, but the diagnostic value of an approach in which the tests and procedures are combined is unknown. The *Aspergillus* antigen galactomannan can be detected in plasma or serum samples from patients with invasive aspergillosis by using a commercially available sandwich ELISA (Platelia *Aspergillus*, Sanofi Diagnostics Pasteur, Marnes-La-Coquette, France) (Stynen *et al.*, 1995). This assay has been found to possess a sensitivity of 67%-100% and a specificity of 81%-98.7% when performed with sera from patients receiving treatment for haematological

malignancies (Verweij *et al.*, 1995; Rohrlich *et al.*, 1996; Sulahian *et al.*, 1996; Poirot *et al.*, 1996; Tabone *et al.*, 1996).

The sandwich ELISA becomes positive at an early stage of infection, and galactomannan has been detected in the sera of some patients even before signs and symptoms consistent with invasive aspergillosis became apparent (Rohrlich *et al.*, 1996). Furthermore, ELISA results are available within 4 hours of collecting the sample (Stynen *et al.*, 1995). Thoracic CT scanning is a major tool for the diagnosis of invasive aspergillosis in neurotropenic patients, and the presence of halo sign or air-crescent sign is highly indicative of disease (Kuhlman *et al.*, 1985). Recently, it was confirmed that the detection of halo-signs in neutropenic patients is an early indicator of the presence of invasive aspergillosis (Caillot *et al.*, 1997). However, the diagnostic sensitivity of a CT scan depends largely on tissue attenuation and an inflammatory response following fungal invasion, so the impaired inflammatory responses of immunocompromised patients may render the interpretation of the results of this technique more difficult than it is when patients are immunocompetent (Rubin & Fischman, 1996).

Radionuclide imaging with ^{111}In-labeled human IgG is both sensitive and specific for detecting microbial invasion, even in the presence of a blunted inflammatory response, since accumulation of ^{111}In-labeled IgG is a result of capillary leakage (Rubin & Fischman, 1996). This technique has been successfully used to detect *Aspergillus fumigatus* infection in granulocytopenic patients (Oyen *et al.*, 1992) and may also be helpful in detecting non respiratory fungal lesions such as those in the sinuses or gut (Rubin & Fischman, 1996).

It is expected that by combining these diagnostic tests and procedures, candidates for treatment will be better selected than they are when the more permissive conventional strategy, which relies solely on the presence of persistent fever that is refractory to antibacterial agents and on the presence of pulmonary infiltrates on the chest roentgenogram, is used. However, since the costs incurred by the alternative diagnostic work-up will be higher, we decided to see whether these costs might be offset by savings made in the overall cost of treatment.

We used decision analysis, a modelling instrument for specifying complex choices in the face of uncertainty, to construct a model for comparing the number of patients expected to receive a diagnosis of invasive aspergillosis and to receive treatment according to each strategy. The expected costs of diagnosis and treatment with DC-Amb were calculated, and the impact of using L-Amb as first-line treatment in at least some cases of probable invasive aspergillosis was assessed. Finally, sensitivity analyses were performed to assess the impact of the necessary assumptions and impact of the estimates of certain variables on the robustness of the conclusions by varying one of the variables in the model while holding the other variables at their baseline value (McNeil & Pauker, 1984).

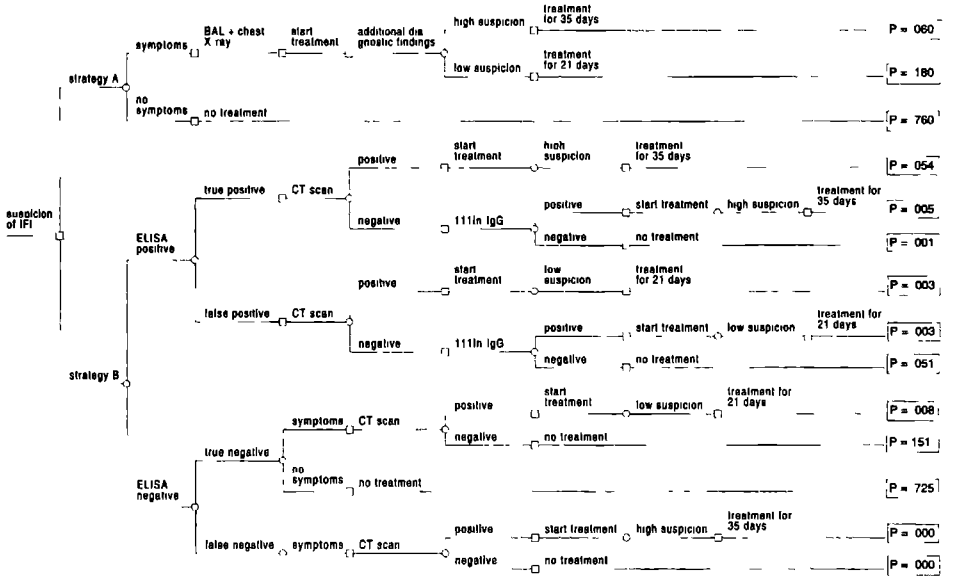


Figure 3.1 Decision tree for two strategies for the management of invasive aspergillosis. The squares in the decision tree indicate a decision node. If one branch continues from a decision node, only one decision is possible. The nodes represented by circles are used if subsequent outcomes occur by chance. Theoretical outcomes of chance nodes with a zero probability, such as no symptoms of aspergillosis after a false-negative ELISA, are not shown in the tree. The path probabilities (*P*) of each branch of the tree are shown. Under baseline assumptions, the probability of receiving antifungal treatment in the conventional strategy (*A*) which is based on persistent fever, bronchoalveolar lavage (BAL), and chest roentgenogram findings, is 0.24 (0.06 + 0.18). The probability of receiving antifungal treatment in the alternative strategy (*B*), based on screening plasma for an *Aspergillus* antigen by using a sandwich ELISA, thoracic CT scanning, and radionuclide imaging (^{111}In-labeled IgG) is 0.073 (0.054 + 0.005 + 0.003 + 0.003 + 0.008 + 0.000).

MATERIALS AND METHODS

Diagnostic work-up

Presently in the Haematology Department at the University Hospital Nijmegen, invasive pulmonary aspergillosis is considered possible diagnosis in neutropenic patients when fever persists for at least 5 days despite treatment with broad-spectrum antibacterial agents and when pulmonary infiltrates are apparent on the chest roentgenogram. Bronchoalveolar lavage is then performed, and treatment is started empirically. Additional chest roentgenograms and culture results are used to determine the level of suspicion of invasive aspergillosis (conventional strategy, Figure 3.1). The alternative diagnostic work-up involves prospective twice-weekly screening of serum or plasma for galactomannan in neutropenic patients

receiving treatment for a haematological malignancy during their stay in the hospital. When galactomannan is detected for the first time, a second sample is obtained for confirmation and, if the antigen is detected again, a high-resolution CT scan of the thorax and sinuses is obtained. If nothing is seen, ^{111}In-IgG scintigraphy is performed.

A diagnosis of invasive aspergillosis will be presumed if antigen is detected and either imaging technique confirms the presence of a pulmonary infiltrate, sinusitis, or another sign of inflammation consistent with invasive aspergillosis. Antifungal treatment will then be started pre-emptively (alternative strategy, Figure 3.1) (Verweij *et al.*, 1996). Since patients may still develop invasive disease despite a negative ELISA result, a high-resolution CT scan will be obtained whenever there are grounds for suspecting invasive aspergillosis.

Table 3.1 Baseline values and ranges of variables and test characteristics used to construct the decision tree for comparing two strategies for managing invasive aspergillosis.

reference	variable	baseline value (%)
hospital data	prevalence of invasive aspergillosis	6
(Rohrlich *et al.*, 1996)	ELISA sensitivity	100
(Rohrlich *et al.*, 1996)	ELISA specificity	94
(Kuhlman *et al.*, 1985)	CT scan sensitivity	89
set	CT scan specificity	95
(Oyen *et al.*, 1992)	^{111}In-labeled IgG scintigraphy sensitivity	90
set	^{111}In-labeled IgG scintigraphy specificity	90
hospital data	patients receiving empirical treatment with desoxycholate formulation of amphotericin B	24

hospital data: data are used from the haematology department of the University Hospital Nijmegen
set: estimates of performance characteristics.

Empirical and pre-emptive treatment.

Currently, 24% of the neutropenic patients in the Haematology Department receive DC-Amb empirically; 18% of these patients receive the drug for an average of 21 days for treatment of possible invasive aspergillosis, whereas the actual prevalence rate of this disease among patients admitted to this department is estimated to be 6%. Since the average course of treatment is 35 days for these patients, we chose this figure to set the duration of pre-emptive treatment for patients whose infections were diagnosed by using the alternative strategy. However, we assumed that patients would receive only a 21-day course of treatment if the

diagnosis of invasive aspergillosis was shown to be unlikely later during the course of treatment.

The probabilities used in the decision analysis were based on our own data and those from the literature (Table 3.1). We relied on estimates of the performance characteristics of the CT scanning and the ^{111}In-IgG scintigraphy published for unselected neutropenic patients, since neither estimate is known for those in whom galactomannan is detected. The costs of diagnosis and treatment with use of both strategies were derived from hospital tariffs, for which diagnostic materials, depreciation costs of medical equipment, administration services, and consultation fees are taken into account (Table 3.2). Treatment costs of both the desoxycholate and lipid formulations of amphotericin B were based on their retail prices, assuming a 70-kg man was given 1 mg/(kg·d) of DC-Amb or 3 mg/(kg·d) of L-Amb, and were estimated to total US dollar (USD) 27 per day and USD 1,115 per day, respectively. No additional costs accruing to toxicity were taken into account because there were no accurate estimates available.

Table 3.2 Costs of diagnostic tests (US dollar, USD) and number of procedures used (N) in the two strategies for the management of invasive aspergillosis.

diagnostic procedure	specification	costs per test (USD)	N
conventional strategy			
chest roentgenogram		44	3
bronchoalveolar lavage	includes cytology and bacteriologic, virological and mycological cultures	562	1
alternative strategy			
aspergillus antigen detection	sandwich ELISA	10	30#
high-resolution CT-scanning	lungs and sinuses	467	1
^{111}In-labeled IgG scintigraphy		555	1

all patients are screened by sandwich ELISA during each neutropenic episode

Decision analysis

The conventional and alternative strategies were defined, and baseline values of the probabilities of each test result, together with the costs of the diagnostic tests and treatments, were estimated and incorporated into the decision tree by using the software program DATA version 2.6 (TreeAge Software, Williamstown, MA). After the path probabilities of each

branch of the tree and the expected costs of diagnosis and treatment per patient were calculated for each strategy, sensitivity analysis was performed.

Primary calculations on the costs of both strategies were based on treatment with DC-Amb to determine whether the additional costs of diagnostic tests included in the alternative strategy were outweighed by any savings resulting from the avoidance of unnecessary treatment. Since different values for both sensitivity and specificity of the sandwich ELISA have been reported (Verweij *et al.*, 1995; Rohrlich *et al.*, 1996; Sulahian *et al.*, 1996; Poirot *et al.*, 1996; Tabone *et al.*, 1996), the impact of varying the performance characteristics of the sandwich ELISA on the positive and negative predictive values and the expected costs of the alternative strategy was determined by varying the sensitivity and specificity with use of values from the literature and the hypothetical values of 50% and 99%, given a 6% prevalence rate of infection. Besides this, we used the variables prevalence of infection (baseline value, 6%), the probability that a patient would receive empirical treatment with DC-Amb (baseline value, 24%), and the costs of the sandwich ELISA (baseline value, USD 10) in the sensitivity analysis.

Treatment with L-Amb

The impact of the availability of L-Amb was evaluated by incorporating three different possibilities of using this expensive drug in the model. First, we assumed that all patients with invasive aspergillosis diagnosed by the alternative strategy would be treated with L-Amb and that all patients whose cases were diagnosed by the conventional strategy would be treated with DC-Amb. Second, we analysed the situation in which all patients with invasive aspergillosis diagnosed by the alternative strategy would be treated with L-Amb and that a varying proportion of those cases diagnosed by the conventional strategy would also be treated with L-Amb. Last, we assumed that the same proportion of patients whose cases were diagnosed according to both strategies would receive L-Amb. In addition, for the second and third scenarios, we calculated the threshold at which the expected costs were equal for both strategies. We also performed a two-way sensitivity analysis to examine the relationship between the proportion of patients treated with L-Amb and the prevalence of invasive aspergillosis.

RESULTS

Probability of receiving treatment

The probability of receiving antifungal treatment as a result of the conventional strategy was 0.24 (Figure 3.1; path probabilities 0.06 + 0.18), while the probability of receiving antifungal

treatment unnecessarily was 0 18 The probability of receiving antifungal treatment as a result of the alternative strategy was 0 073 (Figure 3 1, path probabilities 0 054 + 0 005 + 0 003 + 0 003 + 0 008 + 0 000) with 74% of these cases identified by antigen detection and confirmed by CT scanning The probability of a patient incorrectly receiving treatment was 0 013, whereas only one of every 1,000 patients would incorrectly not receive treatment as a result of a positive ELISA result unconfirmed by imaging The expected costs of diagnosis by the conventional and alternative strategies were estimated respectively to be USD 167 per patient and USD 463 per patient, respectively, and those of treatment were USD 156 per patient and USD 63 per patient, respectively Thus, the calculated savings on therapy of USD 93 per patient, accrued by using the alternative strategy, did not compensate for the extra investment in the diagnostic tests and procedures (USD 296 per patient) Under baseline assumptions, an overall investment of USD 203 per patient would be required to implement the alternative strategy

Table 3 3 Positive predictive value (PPV) and negative predictive values (NPV) and expected costs (US dollar, USD) per patient of an alternative strategy for the management of invasive aspergillosis, calculated for varying values of the sensitivity and specificity of the sandwich ELISA

source of data	sensitivity of ELISA (%)	specificity of ELISA (%)	PPV of the strategy (%)	NPV of the strategy (%)	USD
reference					
Rohrlich *et al* (1996)	100	94	81	99 9	525
Poirot *et al* (1996)	100	89	77	99 9	571
Verweij *et al* (1995)	90	84	73	99 9	615
Sulahian *et al* (1996)	67	94	81	99 8	523
hypothetical values					
	99	99	88	99 9	480
	99	50	54	99 9	922
	50	99	88	99 9	476
	50	50	54	99 7	919

Test characteristics

As expected, the positive predictive value for the conventional strategy was low (25%), and the negative predictive value was high (100%) Irrespective of the level of sensitivity, the positive predictive value of the alternative strategy declined from 88% to 54% as the

specificity of the ELISA decreased, but the negative predictive value remained high (Table 3.3). Decreasing the specificity of the test from 99% to 50% nearly doubled the expected costs per patient because the number of false-positive results increased, leading to the more frequent performance of CT scanning. By contrast, lowering the sensitivity of the ELISA from 99% to 50% had only a marginal effect on the expected costs per patient, reducing them from USD 480 per patient to USD 476 per patient when the specificity was 99%, and from USD 922 per patient to USD 919 per patient when the specificity was 50%. This effect is due to the fact that patients with a false diagnosis of freedom from disease will ultimately develop clinical signs and symptoms consistent with invasive aspergillosis and will still undergo CT scanning.

Sensitivity analyses

When the baseline values for the sensitivity and specificity of the sandwich ELISA were used, the alternative strategy became less expensive than the conventional strategy once the prevalence of invasive aspergillosis exceeded 13% (Figure 3.2), and even when the worst test performance values were used, this threshold increases only slightly to 15.1%. Similarly, the threshold dropped only marginally to a prevalence of 11.5% when the best test performance values of sensitivity and specificity of the sandwich ELISA were used. At a prevalence rate of 13%, the probability of patients having invasive aspergillosis but wrongly not receiving treatment remained approximately one in 1,000. Given baseline values for the performance of the alternative strategy, the conventional strategy became more expensive than the alternative strategy when the likelihood that a patient would be treated with DC-Amb exceeded 48.8%. The expected costs per patient of the two strategies were equal when the cost of the sandwich ELISA was USD 3.30.

Treatment with L-Amb

The expected costs of the alternative strategy were estimated to be USD 3,096 per patient, assuming that each patient with probable invasive aspergillosis was treated with L-Amb. This approach was always be more expensive than the conventional strategy and treatment with DC-Amb, since thresholds could not be identified for the prevalence of infection, for the probability that a patient would receive empirical treatment, or for the costs of the sandwich ELISA. However, if L-Amb was available to all patients, 43.3% of the 24% of patients given empirical antifungal therapy as a result of the conventional strategy could be treated with this formulation before the alternative strategy became less expensive. Assuming that only a proportion of patients were treated with L-Amb, irrespective of the diagnostic strategy used, the one-way sensitivity analysis showed that the expected costs of the alternative strategy were only lower than those of the conventional strategy when >5.3% of all patients given antifungal therapy were treated with L-Amb. The two-way sensitivity analysis (Figure 3.3) showed an

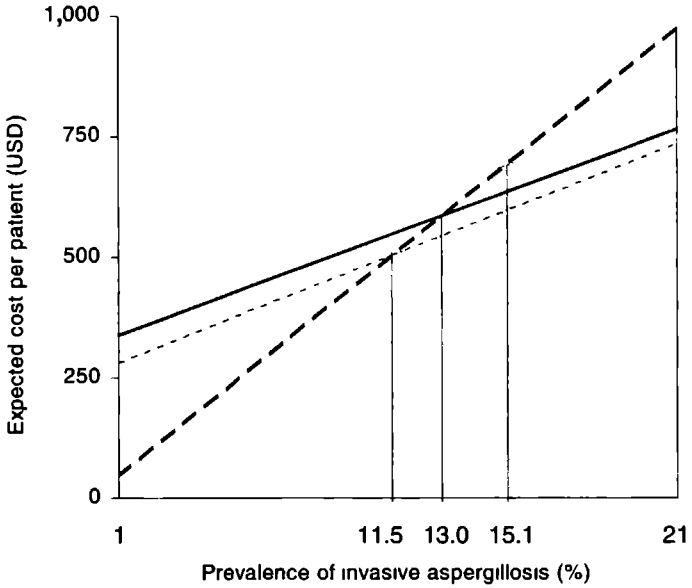


Figure 3.2 One-way sensitivity analysis of the decision tree to determine the effect of the prevalence of invasive aspergillosis on the expected costs per patient. The prevalence of aspergillosis is varied in the sensitivity analysis which results in lines indicating the expected costs of each strategy Where the lines of the strategies intersect (prevalence of 13%, given baseline test performance), this intersection defines a threshold point If the prevalence of invasive aspergillosis is below the threshold, the conventional strategy is optimal, if this prevalence is above the threshold, then the alternative strategy is optimal Baseline values for the sensitivity and specificity of ELISA in the alternative strategy are 100% and 94%, respectively The worst case values, 67% for sensitivity and 84% for specificity, result in a threshold of 11 5%; in the best case these values are both 100%, which results in a threshold of 15.1%. (– – = conventional strategy; —— = alternative strategy, baseline test performance, · · · · · = alternative strategy, worst test performance; – · – = alternative strategy, best test performance).

inverse relationship between the proportion of patients treated with L-Amb for invasive aspergillosis and the prevalence of invasive aspergillosis.

Discussion

The increasing incidence of invasive aspergillosis (Groll *et al.*, 1996), the high morbidity and mortality associated with the infection, and the costs of treatment with novel antifungal agents underscore the need for alternative approaches to managing this disease. The decision tree we constructed can be used as a model to predict the outcome and expected costs of implementing strategies for diagnosing or treating invasive aspergillosis in patients with haematological malignancies We chose the proportion of patients who would receive antifungal treatment, rather than survival, as the measure of outcome since the former variable

is more commonplace and useful, whereas the mortality associated with invasive aspergillosis depends not only on antifungal treatment but also on whether remission from the underlying disease is achieved (Aisner *et al.*, 1977). However, we acknowledge that the baseline values necessary for decision analysis are likely to differ between different institutions and depend on the nature and type of tests and procedures used to diagnose invasive aspergillosis, the prevalence of the disease, and, more important, the guidelines governing the institution of empirical antifungal therapy.

Under the assumed baseline values obtained at our hospital, the probability of receiving treatment unnecessarily could be reduced ~14-fold from 0.18 to 0.013 by using the alternative strategy we propose. We were also encouraged by the finding that the probability of withholding treatment incorrectly was likely to be very low (0.001), although not zero. However, we assumed that no cases would go undiagnosed by the conventional strategy,

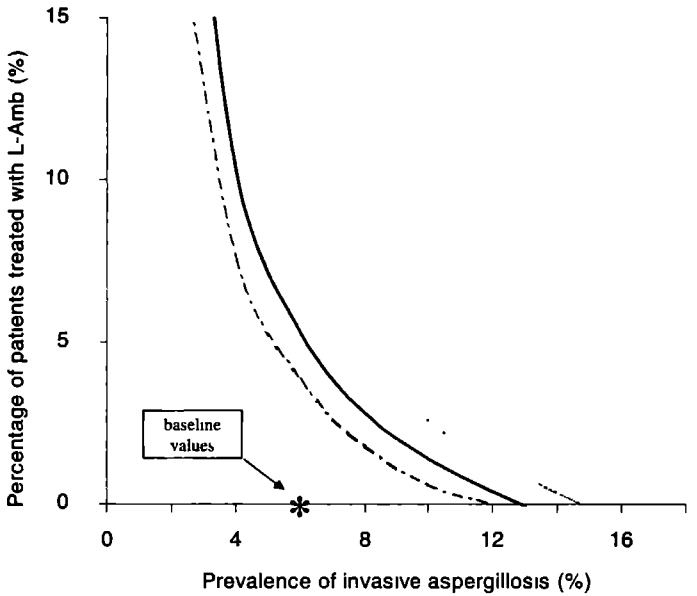


Figure 3.3 Two-way sensitivity analysis of the decision tree to determine the effect of the prevalence of invasive aspergillosis and the percentage of treated patients who received liposomal amphotericin B (L-Amb) in both strategies according to the expected costs per patient As in Figure 3.2, the results for the worst-case and best-case values of sensitivity and specificity, as well as baseline values for sensitivity and specificity of ELISA, are given. The curves indicate the combinations of values for the prevalence of aspergillosis and percentage of patients treated with L-Amb for which the expected costs of the two strategies are equal. Any co-ordinate to the left of the curves favours the conventional strategy, whereas any co-ordinate to the right of the curves favours the alternative strategy. There is an inverse relationship between prevalence of aspergillosis and percentage of patients treated with L-Amb. Thus, with our current prevalence, we would only need to treat one of 20 treated patients with L-Amb, and the alternative strategy would begin to pay for itself. (— : baseline test performance; . . . . . : worst test performance; – . – : best test performance).

which is unlikely to be the case, since invasive aspergillosis only becomes apparent at autopsy for the majority of patients (Groll *et al.*, 1996).

The strategy we propose requires physicians to withhold treatment from patients who would otherwise have received therapy empirically. Typically, these patients would be febrile and neutropenic, but galactomannan would not be detected in blood samples, and CT scans would show not show lesions consistent with invasive aspergillosis. Of course, other fungal pathogens may cause clinical signs and symptoms that are difficult to differentiate from invasive aspergillosis; nevertheless, the fear that patients may die of invasive aspergillosis compels many physicians to begin treatment empirically, after 3-5 days of persistent fever that is refractory to antibacterial therapy, even in the absence of any pulmonary or sinus abnormalities. Thus, as many as two-thirds of all neutropenic patients are given empirical antifungal therapy (Goodman *et al.*, 1992; Winston *et al.*, 1993).

Clearly, the withholding of treatment will be difficult to achieve in such circumstances because the morbidity associated with DC-Amb therapy is regarded as less harmful than is the withholding of treatment from a patient with disease. Furthermore, when other less toxic antifungal agents are made available, the threshold for treating patients empirically will be lower still, with only the high cost of novel antifungal agents and the constraint of resources providing disincentives to using these drugs empirically. The evaluation of alternative strategies for managing invasive aspergillosis is therefore imperative if we are to define the place of new antifungal agents and control their use.

The baseline values for the performance of the sandwich ELISA were based on the results of a prospective study of children during treatment for a haematological malignancy (Rohrlich *et al.*, 1996). However, false-positive ELISA results have been reported to occur within 30 days after bone marrow transplantation (Sulahian *et al.*, 1996). and within 10 days after cytotoxic therapy (Swanink *et al.*, 1997). In the latter study, antigen was detected intermittently in a series of serum samples from some patients who had no evidence of invasive aspergillosis, which necessarily lowers the specificity of the sandwich ELISA (Swanink *et al.*, 1997). However, since sequential tests and procedures are used in the alternative strategy to diagnose invasive aspergillosis, the impact of a less sensitive and specific ELISA on the positive and negative predictive values of the strategy as a whole is likely to be negligible. Nevertheless, the total costs per patient will increase considerably, especially if the specificity proves low. In addition to better selection of patients for antifungal treatment, the use of the sandwich ELISA as a screening test for galactomannan may also allow for the diagnosis to be made at an early stage of infection, at least for some patients (Hiemenz *et al.*, 1995; Stynen *et al.*, 1995; Verweij *et al.*, 1995).

Under baseline assumptions, we found that the current conventional strategy was less expensive per patient than was the alternative strategy when patients were treated exclusively with DC-Amb. However, we omitted the costs related to toxicity of treatment with this formulation even though they were expected to be high because a larger number of cases diagnosed by the conventional strategy will receive treatment. As a consequence, all

calculated expected costs were a conservative estimate of the actual costs. Nevertheless, the alternative strategy we propose will become less expensive than the conventional strategy if >5.3% of the treated patients are treated with L-Amb and it will become even less expensive if it proves tenable in circumstances where the majority of patients with haematological malignancies currently receive empirical antifungal treatment.

Moreover, the probability of invasive aspergillosis is also high among these patients, so those cases of invasive aspergillosis that are diagnosed by using the alternative strategy may benefit the most from being given first-line treatment with L-Amb. However, the costs of treating all such patients with L-Amb will be high; thus a comparative study of L-Amb and DC-Amb is still required to determine the feasibility of this approach by taking into account other important aspects such as outcome, adverse drug reactions, and costs.

In conclusion, we constructed a decision model to explore the feasibility of implementing an alternative strategy for diagnosing invasive aspergillosis in patients receiving treatment for haematological malignancies. This model indicated that screening for galactomannan and using imaging techniques, as compared with the conventional strategy, will reduce the number of patients who require treatment and therefore may help to control the use of toxic or expensive antifungal agents. The model showed the relevance of defining a time horizon when evaluating diagnostic technologies. Limiting the time horizon to the diagnostic process itself showed that the financial implications of using ELISA for screening of serum for detecting galactomannan and possibly obtaining a CT scan and a ^{111}In-IgG scintigraphy is more costly and equal effective in terms of detecting cases compared to the conventional diagnostic strategy. However, expanding the time horizon in this model, thus including the treatment process leads to the conclusion that the financial investment needed for the alternative diagnostic work up can be compensated by less costs of treatment in the case of using the newly developed lipid formulations of amphotericin B.

CHAPTER 3.2
THE ISSUE OF THE RELEVANT COSTS AND CONSEQUENCES: DETERMINING THE PERSPECTIVE OF THE ANALYSIS

Based on Severens JL, Brokx JPL & Broek P van den (1997). Cost analysis of cochlear implants in deaf children in The Netherlands. *American Journal of Otology* 18: 714-718.

INTRODUCTION

The issue of relevant costs and consequences, in particular the role of determining the perspective of analysis, is important when performing economic evaluation alongside clinical trials (Drummond & Davies, 1991). The definition of the perspective influences the relevancy of the different costs that have to be analysed. Of course, the choice for the units in which costs are analysed is of influence of the final results of a cost analysis. Besides this, the choice of the perspective determines the way cost prices should be calculated. Therefore, when comparing results of cost analyses of different studies, comparison of the results solely is hardly useful. The methods used for measurement of costs should be examined before conclusion about discrepancies can be made. This issue is illustrated by a study on costs of cochlear implants, which included comparison with the results of several studies found in the literature.

On the basis of a series of studies demonstrating safety and efficacy, multichannel cochlear implants were approved by the Food and Drug Administration for use in adults in 1984 and in children in 1990 (Balkany, 1993). Up till 1995, approximately 6000 cochlear implant procedures had been undertaken world-wide (Lea & Hailey, 1995). There is no comparable alternative medical treatment for total deafness. In many countries policy makers are faced with the decision whether or not to include cochlear implants to the basic benefit package. In the face of scarce resources, policy makers and health care purchasers are not only interested in the effects of certain health care interventions, but also in the costs, which are involved. Several studies have been published which report about costs of cochlear implants, which cover only a few countries: Australia (Lea, 1991; Lea & Hailey, 1995), the United Kingdom (Summerfield & Marshall, 1995a; Davis *et al.*, 1995; Hutton *et al.*, 1995) and the United States (Wyatt *et al.*, 1995; Harris *et al.*, 1995). Some of these include analysis on implantation programs for children. However, differences in health care settings in the different countries influence the results of a cost analysis (Summerfield *et al.*, 1995b). This makes application of the results in a specific policy making context difficult. This article describes the results of a cost analysis, which was performed alongside a clinical study of cochlear implants in children in The Netherlands. The results regarding the effectiveness of cochlear implants in children analysed in this study are reported elsewhere (Vermeulen *et al.*, 1995; Snik *et al.*, 1997a; Snik *et al.*, 1997b; Coerts *et al.*, 1996).

METHODS

Between 1993 and 1996 106 prelingually deaf children were screened as candidates for a cochlear implant. Of these, 20 children were implanted. An extensive description of selection, inclusion and exclusion criteria is reported elsewhere (Vermeulen *et al.*, 1995). The children

received a multichannel cochlear implant, the Nucleus Mini System 22 with a MSP Processor. Their mean age was 7 years, 1 month (range 3;11 to 11;11).

A societal perspective was used for the analysis. This perspective implies that real costs of medical care were calculated, instead of using tariffs. In addition, non-medical costs, like patients or parents travelling costs were included in the analysis. The time horizon was five years. The cost analysis of selection, implantation and rehabilitation during the first year were based on empirical data. Costs that are incurred during the remaining years, were based on the planned after care. Costs of maintenance of the external cochlear implant hardware during this period were estimated.

The data used for the costs analysis were obtained from two institutes involved in the project. For the selection and implantation data from the University Hospital Nijmegen were used, based on respectively 106 and 20 children. In one of the 20 children use of the cochlear implant was abandoned after one year because inadequate hearing sensation levels. This child had a partial insertion because of total obliteration of the cochlea after meningitis. The data of the remaining 19 children were used for the cost analysis of rehabilitation and after care. These data were obtained from the Institute for the Deaf in Sint Michielsgestel.

The measurement of volumes

Volumes of utilisation of human resources and materials were prospectively registered. Time spent by different types of personnel was recorded during the phases of selection, implantation and rehabilitation. Utilisation of facilities such as operation theatre, recovery room, audiological centre and the special rehabilitation centre was recorded in production entities like number of hours operating time, number of audiology contacts and days in the rehabilitation centre. Besides this, hospitals days, out patient visits and return visits for rehabilitation were registered. Registration covered one year follow up after the implantation of each child. For the subsequent period volumes were modelled on the basis of planned after care activities.

The measurement of prices

Basis for the calculations were 1994 prices[7]. If prices were not available for this year, a price index was used to make the necessary inflation corrections. Overhead costs of general departments like hospital administration and personnel department were not included. Prices of the different personnel categories involved (ENT specialist, audiologists, psychologist, speech therapist etc.) were based on the midpoint on the scale for each grade of professional

7. Prices are mentioned in Dutch guilders (Dfl.). Mean 1994 exchange rate for Dfl. 1 is 0.55 US dollar.

involved, including scale specific social taxes. Expenses for other salary overheads like holiday premiums and fringe benefits were added (8% and 3% respectively).
Costs of materials used were based on retail prices. Capital costs for this equipment were based on costs of depreciation, interest and a surcharge for annual maintenance costs (8%) (Rutten *et al.*, 1993). Depreciation and interest were based on annuities of the initial capital outlay and the economic lifetime of the equipment involved. The annual annuity and maintenance costs were divided by the annual production number. For instance, annual costs for general equipment in the operating theatre were divided by the annual number of operating hours. The annual costs of operating equipment specific for cochlear implants were divided by the number of implants performed annually. Costs for using accommodation were added in proportion to the time that the accommodation was used. Energy costs and cleaning costs per square meter were calculated on the basis of actual space used. The price of a hospital day was obtained by dividing the annual costs for nursing staff, materials used, feeding and other hotel costs by the number of hospitals days realised in the ENT department. The calculations resulted in Dfl. 679 per hospital day. The price per hour of the ENT out patient department was determined on the same basis, which resulted in Dfl. 102 per hour out patient department time (ENT specialist excluded). Diagnostic tests like CT scan, MRI and ECoG/EBER were valued according to the appropriate tariffs. Reason for not performing extensive cost analyses was that these tariffs were an approximation of the integral price (Rutten *et al.*, 1993). For rehabilitation, a special cochlear implantation centre was available which was not used for any other purpose. The annual costs of this facility were calculated on an integral basis, including annual costs of the building, equipment, power etc. These annual costs were divided by the number of patients that entered the rehabilitation phase annually. Costs of after care which were planned to take place after the year of implantation were discounted on the basis of a 5% discount rate (Drummond *et al.*, 1987).

Non medical costs consisted of travelling costs of the children and parents. On the basis of the postal codes travelling distances for the children and parents to the institutions were estimated by means of a route-planning program. In accordance with Dutch guidelines for cost analysis in health care a price of Dfl. 0.39 per kilometre was used (Rutten *et al.*, 1993).

Sensitivity analysis

To assess the impact of certain variables on the robustness of the conclusions, a sensitivity analysis was performed. During the clinical study 20 out of 106 children (19%) were selected for having a cochlear implant. This rate of implanted children as part of the number of screened children was varied between 9 and 29% as being possibly important on the estimated total costs of cochlear implant per child.

RESULTS

Costs of selection

106 deaf children entered the selection phase for a cochlear implant. The application and intake of each of these children consumed relatively little time of the members of the cochlear implant team. Surcharges for accommodation and costs of used administrative materials were added. This resulted in Dfl. 1,228 per child, and Dfl. 130,154 in total.

The screening resulted in 20 candidates who were considered suitable for a cochlear implant. These children were subjected to audiological, psychological and medical tests. All of them underwent a CT scan to see if the cochlea was suited to insert the electrode arrays. For seven of the children general anaesthesia was needed to perform the CT scan, which required a one-day stay in the hospital. More costly were tests like MRI (sometimes with general anaesthesia and thus one day hospital stay) and ECoG/EBER. The latter was done operatively under general anaesthesia and required a three day hospital stay. However, these tests were only applied to five of the children. The costs for audiological, psychological and medical testing were Dfl. 94,349 total.

Besides these child-specific activities, costs of the team involved in this selection phase were analysed. These activities consisted mainly of team meetings to discuss candidacy of the children for a cochlear implant. Besides travelling costs, main part of this was the actual time spent by the team members, which resulted in Dfl. 60,608 in total.

Thus, the total costs involved of the selection procedures, amounted Dfl. 285,111. These costs were ascribed to the 20 implanted children. This resulted in Dfl. 14,256 selection costs per implanted child.

Costs of implantation

Main costs of the implantation phase were the costs of the cochlear implant hardware. Although the internal hardware is meant to last a lifetime and the external hardware for at least eight years, no depreciation calculations were performed. Reason for the strategy to consider the price of the hardware as costs in the year of implantation, was that the cochlear implant is solely used for the benefit of one patient. At the time of the study the price for the hardware was Dfl. 46,397. Small materials used, including costs of sterilisation were Dfl. 2,098 per operation. Capital costs for medical equipment was calculated as a surcharge for an implant operation. Implant specific equipment resulted in Dfl. 58, ENT specific equipment in Dfl. 107 and general equipment in Dfl. 298 per operation. Including overhead costs of the operating theatre (cleaning, housing) this resulted in Dfl. 592 overhead costs per operation.

The time spent in the operating theatre formed the basis for calculating the costs of personnel, which resulted in Dfl. 2,281 per operation. In total the implantation costs were Dfl. 51,368 per child.

The costs for a one-day stay in hospital were determined to be Dfl. 679. Mean stay in hospital was 6.6 days. After a patient had been discharged an out patient visit was planned to check the recovery of the operation wound and overall condition of the patient. This visit lasted 45 minutes, costing ENT specialist time and a surcharge for accommodation (total Dfl. 162). The costs for hospital stay and out patient clinic visit resulted in Dfl. 4,646 per patient. Together with the cost of the operation itself, the costs of the implantation were Dfl. 56,014 per child.

Costs of rehabilitation

All activities in the rehabilitation phase took place in the cochlear implant rehabilitation centre of the Institute of the Deaf in Sint Michielsgestel. Per rehabilitated person Dfl. 4,947 for infrastructure costs was calculated. The remaining costs were calculated on a variable basis. Not only the institute's audiologist, audiology assistant, speech pathologist and speech therapist were involved in this phase, but also the child's teacher and the school speech therapist. The rehabilitation phase can be divided in several stages. Several weeks after implantation, the external cochlear implant hardware had to be fitted. The time spent by the different persons involved was 28 hours by the audiologist, 4 hours by the audiological assistant and 20 hours by the speech therapist. This time spent incurred Dfl. 3,110 per child. After the regulation, the actual rehabilitation started. A child would stay with one or two parents in the centre for two times a midweek. Each day several sessions were performed with a child. For efficiency reasons, two implanted children stayed in the centre at the same time. Analyses of time spent by the team members resulted in 6 hours audiologist time, 201 hours speech therapist time and 16 hours speech pathologist and school speech therapist time. For each rehabilitated child this resulted in estimated costs of Dfl. 13,611. After this period of intensive rehabilitation, the child's progress is assessed regularly. This was done either at the cochlear implant rehabilitation centre or at the regular school. These assessments were done monthly during the first three months, once every six weeks during the remaining months of the first year. This assessment period resulted in Dfl. 3,039 per child and existed only out of costs for personnel. In total the costs of the rehabilitation phase were Dfl. 24,707 per child.

Costs of long term care

During the second year after transplantation, assessment and tuning days are planned four times, during the third year two and during each subsequent year 1 contact day is planned. Each visit consists of several hours of contact between the patient and the different team members involved. The total costs for a day were calculated to be Dfl. 1,556. Besides this,

each child receives a fixed number of hours speech and hearing therapy annually, which results in Dfl 1,093 per child per year Adding up the different discounted costs per year, this results in Dfl 15,249 per child for this four-year post-operative period

Besides the costs of the after care days, maintenance of the cochlear implant hardware was estimated The external hardware needs periodic maintenance and has more breakdowns in children than in adults For this reason one spare processor is needed per eight children Besides this the infrastructure for maintenance and small replacement materials were needed,

Table 3 4 Costs (in US Dollars) of cochlear implants in children

selection				
	cost per child	number of children	total costs	cost per implanted child
application and intake	1,228	106	130,154	
psychological and medical test	4,717	20	94,349	
team activities			60,608	
				14,256
implantation				
			costs per child	
hardware			46,397	
operation costs			4,971	
hospital days			4,484	
out patient visit			162	
				56,104
rehabilitation				
			costs per child	
regulating hardware			3,110	
rehabilitation			13,611	
assessment			3,039	
overhead costs			4,947	
				24,706
long term care (until 5th year)				
			costs per child	
contact days			15,249	
maintenance			7,393	
				22,641
total medical costs per implanted child				117,617
non-medical costs per child				3,838

which leads to the estimation of Dfl. 2,085 per child per year. The discounted maintenance costs for the five years is Dfl. 7,393 per child

Adding all after care costs per year and using a discount rate of 5% the total costs for the after care for the second until the fifth year is Dfl. 22,641 per child.

Non medical costs

Non medical costs related to the cochlear implant procedure consisted of travelling costs of the parents accompanying the child. The children came from all over The Netherlands, with a mean distance of 133 kilometre to the University Hospital and 139 kilometre to the Institute for the Deaf. The number of visits times the costs for travelling resulted in mean travelling costs for the whole period of five years. Like the selection costs, the travelling costs of all the children who entered the selection phase were ascribed to the finally implanted children. Using a 5% discount rate for the costs after the first year, the travelling costs were Dfl. 3,838 per implanted child.

Overall result

The overall results of the cost analysis of cochlear implants in children are presented in Table 3.4. The total medical costs per implanted child are Dfl. 117,617, given the baseline rate 19% of implanted children as part of the number of screened children. Varying this rate in a sensitivity analysis between 9 and 29% results in total costs per child of Dfl. 129,274 and Dfl. 114,232 respectively, which means that the rate does not have a large impact.

Discussion

The result of our cost analysis of cochlear implants in children is primarily useful for reimbursement issues of policy makers. No comparison has been made with another facility since the clinical study was a non-comparative observational study. For this reason, the concept of opportunity costs, e.g. financial comparison with alternative facilities for the deaf, has not been applied.

The costs related to the educational situation of the children were not part of the cost analysis, due to the limited period of follow up. However, Hutton *et al.* (1995) suggested that the impact of cochlear implant on education is a key factor on the evaluation of cochlear implants in children. They presented very different estimates. Considerable lifetime savings on costs of education have been estimated (£ 51,265, approximately Dfl. 133,389) but as the authors state these results must be treated with caution. Other longer-term financial benefits might occur when persons involved are prevented from being dependent on social services and could contribute to taxation instead of draining from the social security system (Roberts,

1993). Concerning cochlear implants in adults an increase in income after implantation was measured (Harris *et al.*, 1995). However, as long as the impact of cochlear implant on the educational setting of the children is not properly investigated it will remain difficult to predict any changes in costs for education and eventually their employment status (Moog & Geers, 1995).

The societal viewpoint of the analysis would require the incorporation of more relevant cost categories in our analysis. Parents spend considerable time accompanying their deaf child during the selection and implantation phase. Besides this, especially the rehabilitation and after care phase require parents' time. The indirect costs involved (lost labour time) were not part of our analysis due to the fact that the relevant data were not available.

During our study no major complications with the 19 children were faced. Major complications could cause our calculations to be a serious underestimation. In the literature only few complications have been reported (National Agency for Medical Development and Evaluation, 1994). Removal of the implant was only necessary in 0.6 to 1% of the cases. Besides this, complications like anaesthetic complications, flap related problems, and electrode placement problems might occur. Of course such an event would increase costs for treatment for the specific patient. However, regarding the very small chance, the mean costs of cochlear implants in children do not increase significantly. The results of sensitivity analyses performed in published studies showed that the influence of these probabilities was negligible on the ratio between costs and quality adjusted life years gained (QALY) (Summerfield & Marshall, 1995a; Wyatt *et al.*, 1995).

The price of the implant hardware is a large part of the total costs of cochlear implant procedure as is confirmed by our study. Other studies show that the ratio between costs and effects is highly sensitive to this price. The price might fall in future (Summerfield *et al.*, 1995a), however, no considerable change in price of the cochlear implant hardware can be expected for the time being (Roberts, 1993).

Comparison of results

Compared to the results from cost analyses in other countries, the costs of the paediatric cochlear implants program in the Netherlands are relatively high. Most differences however, can be explained by methodological differences.

In a decision modelling approach, costs of cochlear implantation in the UK were described (Hutton *et al.*, 1995). The costs of selection were £ 790 (approximately Dfl. 2,000), costs of implantation were estimated to be £ 15,522 (Dfl. 40,237) and the rehabilitation phase is estimated to cost only £ 900 for the first year and £ 3,750 for the other years (Dfl. 12,090 in total) until 16 years of age. Although the costs were categorised in the same phases as in our study, straightforward comparison of the results is difficult since no information was provided concerning the sources for the cost data, overhead and accommodation costs and the year of

the prices. Besides this, the costs of the selection phase were not ascribed to the children who were implanted. As the authors state, the results of their exploratory work must be treated with caution because the analyses incorporate a very large number of assumptions.

The studies by Lea (1991) and Lea & Hailey (1995) mention costs for selection, operation, the implant hardware ($17,030, equals approximately Dfl. 30,995) and rehabilitation for cochlear implants for prelingually children, which in total would be $36,630. However, these costs were based on Australian 1991 fees instead of real costs, which invalidates comparison with the results of this study. Besides this, relatively low prices for the hardware were used in the calculations.

The estimated costs mentioned by Wyatt *et al.* (1995) were based on a decision analytic model which concerns postlingually adults. Because of the essential differences in the selection and rehabilitation phase, only the costs of the implantation phase can be compared with our results. Total costs of implantation consist of the implant hardware costs ($19,383, approximately Dfl. 35,277) and the operating costs ($12,227, approximately Dfl. 22,253). It is not clear whether these costs were based on fees or on real costs. However, concerning the operating costs these results were considerable higher than our results.

In the studies by Davis *et al.* (1995) and Harris *et al.* (1995) only costs of the implant hardware and in the latter study, costs of operation were considered. The basis for these estimations is not clear. Considering our results these studies seem to underestimate the costs that were really involved in cochlear implants.

Because costs of cochlear implants involved in adults can not be translated to children, the study by Summerfield & Marshal (1995a) is relevant. In this study the costs of a children's implant program in the United Kingdom is analysed. The authors find total costs at the end of the first year to be £ 24,250 (approximately Dfl. 63,050). Taking into account the much lower price for the implant hardware used in their estimates and the general differences in the paediatric programs, their result is in line with our calculations.

ACKNOWLEDGEMENTS

Mr. J. Cornelisse is greatly acknowledged for providing the necessary data. The authors benefited from the comments of G.J. van der Wilt, Department of Medical Technology Assessment, University of Nijmegen. This research was made possible by a grant from the Reinier Post Foundation.

CHAPTER 4
THE ISSUE OF THE ACCURATE MEASUREMENT OF COSTS AND CONSEQUENCES: INCORPORATING PRODUCTIVITY COSTS

Based on Severens JL, Mulder J, Laheij, RJF & Verbeek ALM. Precision and accuracy in measuring absence from work as a basis for calculating productivity costs. *Social Science and Medicine* [accepted for publication].

INTRODUCTION

Regarding the accurate measurement of costs and consequences, one of the topics about which debate is going on among health economic reseachers, is the measurement and valuing of productivity loss. Productivity loss is related to the consequences of persons' inability to work due to illness. The main point of the discussion concentrates on whether to include the consequences of the inability to work in monetary terms in de numerator of the ratio of costs and effectiveness or whether these consequences are to be measured in utility terms as an outcome measure and therefore must be included in the denominator of the ratio (Brouwer *et al.*, 1997a; Weinstein *et al.*, 1997; Brouwer *et al.*, 1997b). However, in cost-effectiveness analyses that do not use utility measures for effectiveness it seems obvious to reflect the consequences of inability to work in monetary terms in the numerator of the cost-effectiveness ratio, thus valuing productivity loss as being costs (Luce *et al.*, 1996).

When analysing productivity costs, a distinction can be made between lost productivity related to absence from paid work, reduced productivity at paid work, and lost home productivity (van Roijen *et al.*, 1996). Although the latter two can be of importance, in our study we concentrate on productivity costs related to absence from paid work. When analysing productivity costs due to absence from work, days absent from work are to be measured, after which the number of days is valued (Koopmanschap & Rutten, 1996b).

To determine absence from work due to illness several instruments can be used. Sick leave registers would be a reliable source of information to obtain the number of days sick leave for participants in a study. However, when analysing the productivity cost of study participants related to a specific disease who, of course, are employed in different localities or companies this approach is not practical. To overcome this problem questionnaires are used to measure absence due to illness to be able to calculate productivity costs. Such questionnaires can be applied in a more or less prospective manner, but more often data are gathered in a retrospective way. Different recall periods to measure absence from work can be found in the literature and to our information the longest period used was 12 months (Bertera, 1991; Agius *et al.*, 1994; Jones *et al.*, 1995). The question arises whether using such questionnaires leads to valid results because a potential for recall bias exists in every study in which historical self-reported information from respondents is used. The imperfect memory of respondents can harm the precision (difference by chance between memory and fact) and accuracy (systematic difference between memory and fact) of the sick leave data, which, theoretically, can cause recall bias. This can influence the absolute level of costs induced by absence from work. However, recall bias is only relevant when accuracy of recall regarding the measure of interest is different between the different groups, which can be distinguished in the study (Raphael, 1987).

The purpose of this study was to study precision and accuracy of a retrospective self-administered questionnaire on sick leave by comparing the self-reported absence with company-registered absence data. Different recall periods were used to analyse the relationship between memory and length of the recall period.

METHODS

A local branch of a pharmaceutical company involved in research, marketing, and sales participated in the study. All employees, existing largely of office workers, were asked to fill-in a questionnaire to indicate the number of days absent from work due to illness concerning five different recall periods: the past 2 weeks, 4 weeks, 2 months, 6 months, and 12 months. The actual date the questionnaire was filled-in was written on the questionnaire by the respondent. The dates of the five different recall periods were defined separately for each respondent. The company kept a prospective sick leave register. Using this register for each respondent true absence from work was determined for the five recall periods. These register-based data were considered to be the gold standard regarding absence due to illness. Data on sick leave for non-respondents were not made available.

Mean and standard deviations of company-registered and reported sick leave were calculated for the subsequent recall periods for both all respondents and cases only. The latter were defined as respondents with reported or registered absence. The mean difference was calculated based on each respondent's difference in registered and reported number of days absent from work due to illness for the various recall periods. Accuracy was studied as a possible systematic difference between the data and was tested by the sign rank test, which tests if the differences calculated between registered and reported have a symmetric distribution with a zero mean. These analyses were performed for both all respondents and cases only.

Analysing the precision of retrospective measurement of sick leave, the percentage of respondents without any difference between registered and reported sick leave was determined. Besides this perfect precision several other levels of precision were introduced. These levels were defined as the maximum absolute discrepancy between registered and reported sick leave using the categories 1, 2, 3, 4, 5, 6, 7, 8, and 9 or more days discrepancy. For each level of precision, the cumulative percentage of the respondents with perfect agreement or a discrepancy was determined and plotted for the different recall periods. The same calculations were performed only using the data of the cases.

For the respondents with registered absence from work, the question was addressed how the relative difference between registered absence form work and reported absence from work changed regarding the length of the recall period. Thus, bias was reformulated as a proportion between registered and reported absence from work: ((registry–reporting)/registry). The mean and median proportions were calculated for the different recall periods.

RESULTS

At the time of this project 210 people worked at the company; all were asked to participate in the study. Of these, 155 returned the questionnaire (response rate 74%). One questionnaire was incomplete, and the reported results could not be matched to the data from the company registry, resulting in a study sample of 154. It was possible to make company-registered absence due to illness available for all respondents except one. Reported sick leave for the different recall periods was limited slightly due to missing data. The agreement between reported data and registered data was based on the respondents for whom both company-registered and reported data were available.

Table 4.1 Registered absence from work due to illness in days (number (N) of respondents, mean, and standard deviation (s.d.)), reported absence from work in days (number (N), mean, and standard deviation (s.d)), and difference between registered and reported absence from work in days (number (N), mean, min to max, and p-value (p) of the sign rank test) regarding the different recall periods for all respondents, and cases (respondents with reported or registered absence).

	registered absence			reported absence			difference (registered minus reported absence)			
	N	mean	s.d.	N	mean	s.d.	N	mean	min to max	p
2 week recall										
all respondents	154	0.1	0.7	146	0.1	0.6	145	0.03	0 to 1	0.06
cases	10	2.3	1.6	10	1.8	1.6	10	0.50	0 to 1	0.06
4 week recall										
all respondents	154	0.3	4.3	144	0.2	0.8	143	0.06	-1 to 3	0.09
cases	15	2.6	1.8	14	2.0	1.9	14	0.57	-1 to 3	0.09
2 month recall										
all respondents	154	0.7	1.9	143	0.4	1.3	142	0.32	-5 to 10	0.00
cases	31	3.6	2.6	27	1.8	2.5	27	1.70	-5 to 10	0.00
6 month recall										
all respondents	154	2.4	4.4	143	2.1	4.5	142	-0 01	-31 to 18	0.37
cases	81	4.5	5.3	74	4.1	5.6	74	-0.03	-31 to 18	0.37
12 month recall										
all respondents	154	4.6	10.3	145	3.9	8.3	144	0.33	-40 to 46	0.76
cases	97	7.2	12.2	90	6.3	9.8	89	0.53	-40 to 32	0.76

In Table 4.1 the results of days absent due to illness are reported regarding both all respondents and cases. As expected, the number of cases increases with the length of the recall period. The longer the recall period, the higher the number of days sick leave, both for

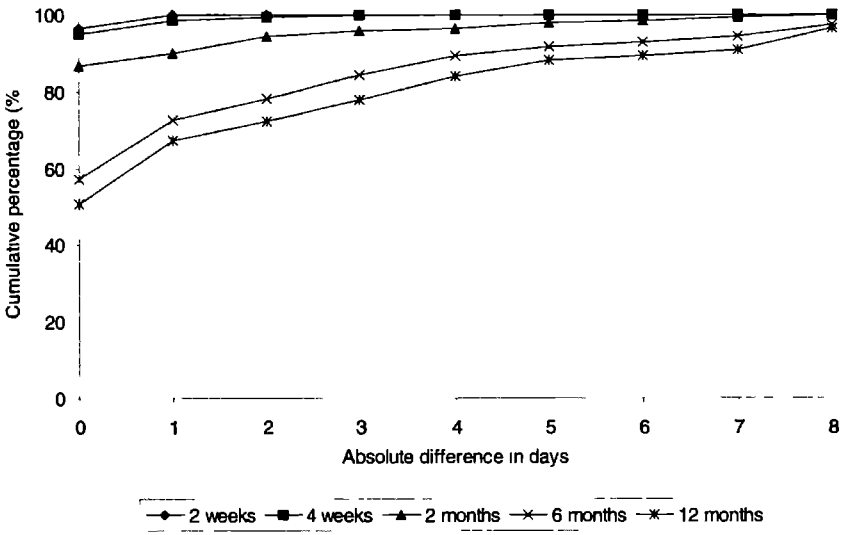


Figure 4.1 Cumulative percentage of all respondents with zero to a maximum of 8 days absolute difference between reported number of days of sick leave, and registered number of days of sick leave for five different recall periods.

all respondents and cases. The comparison of company registered sick leave and respondent reported sick leave is reflected in an absolute difference for each respondent for the various recall periods of which the mean is mentioned. The results of the sign rank test for both all respondents and cases showed no systematic positive (underestimation of number of days absence) or negative (overestimation of number of days absence) deviation between registered and reported sick leave.

The cumulative percentages of respondents who did not reveal any difference between registered and reported sick leave, and an absolute difference of a maximum of respectively 1, 2, 3, 4, 5, 6, 7, 8, and 9 or more days for the different recall periods are shown in Figure 4.1. As can be seen from this figure more than 95% of the reported data matched the registered data perfectly when the recall period was limited to 2 weeks, and 4 weeks. This percentage decreased to 87%, 57%, and 51% for respectively the recall periods 2 months, 6 months, and 12 months, respectively. Accepting, for example, a margin of a maximum of 3 days difference between registered and reported data, the percentages of no difference for the five recall periods were 100%, 100%, 96%, 85%, and 78%, respectively.

When selecting the cases, the results are different (Figure 4.2). As mentioned in Table 4.1 the number of respondents included in this analysis was limited, respectively 10, 15, 31, 81, and 97 for the subsequent recall periods. For these subgroups a discrepancy between registered and reported absence exists for 50% of the respondents regarding even the shortest recall

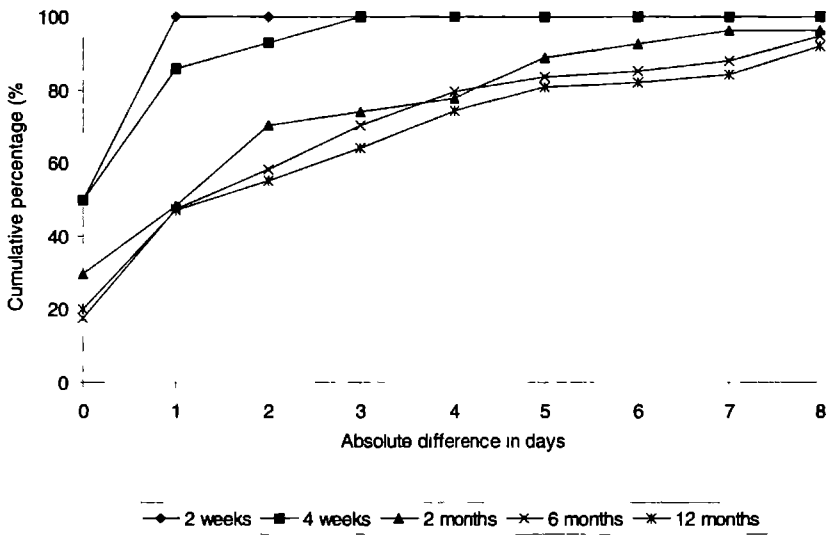


Figure 4.2 Cumulative percentage of respondents with registered or reported absence from work with zero to a maximum of 8 days absolute difference between reported number of days sick leave, and registered number of days of sick leave for five different recall periods.

period (2 weeks). For the longer recall periods this figure gets worse (up to 80% for 6 months, and 12 months). Where a 3-days discrepancy is regarded acceptable, this results in 100%, 100%, 74%, 70%, and 64% agreement between registered and reported data.

The mean and median proportion for the different recall periods were respectively 34% and 10%, 30% and 0%, 46% and 50%, 7% and 30%, and –16% and 0%.

Discussion

The issue of the accurate measurement of costs and consequences was illustrated by the concept of productivity costs. To determine these costs, when performing economic evaluations alongside clinical trials, measurement of absence from work is inevitable. For reasons of reliability, prospective registration of absence from work is to be preferred above retrospective analysis. However, this indicates a substantial workload for participants, which can be one of the causes of missing data (Goossens *et al.*, 1998). Using only a short period of prospective registration, and multiplying the results afterwards to estimate absence from work for a longer term leads to valid results only when large numbers of respondents are included in a study (van Roijen *et al.*, 1996). Also, in this manner an increase or decrease in absence from work during a longer period can not be measured. Up until now recall periods to measure

absence due to illness have been established rather arbitrarily. Burdorf *et al.* (1996) state that a 6-month recall period was chosen to avoid strong recall bias. Other studies chose, again arbitrarily, a recall period of 12 months (Bertera, 1991; Agius *et al.*, 1994; Jones *et al.*, 1995).

The purpose of our study was to study precision and accuracy of a retrospective self-administered questionnaire on sick leave by comparing the reported results with company-registered absence data for different recall periods. Although in our study the number of cases was limited, the results suggest that accuracy seems to be acceptable when measuring sick leave retrospectively as a systematic over- or underestimation could not be detected. When analysing the precision to measure absence due to illness in a retrospective manner, researchers should be aware of a certain imprecision which increases when a longer recall period is used. To choose a recall period and therefore introducing a certain level of imprecision to our opinion should be based on the purpose of the study. Suppose a situation in which a health care intervention is evaluated of which medical costs are expected to be more or less the same compared to productivity cost; when the latter are calculated on the basis of reported sick leave, the precision level acquired would be rather high. In this case, imprecision of measuring sick leave might influence the conclusions of the evaluation.

The mean and median proportion ((registry-reporting)/registry) for the different recall periods varied across a wide range. Because a clear function between recall period and proportion was not found, for instance increasing proportional errors, a systematic bias seems not to be the case. Besides this, in the context of the analysis of productivity costs, which should report absolute cost numbers, the meaning of the proportions is limited. A mean proportion of 34% when using a 2 week recall period (normally about 10 working days) can only be an absolute number of 3.5 days maximum, while a mean proportion of -16% (which seems to be more accurate) when using a recall period of 12 months (normally about 210 working days) can be as high as -33.6 days.

The overlapping recall periods or time frames defined in the questionnaire were also used in the analysis. The possibility existed that the responses to the questions regarding the different recall periods correlated and therefore the data of the adjacent periods in stead of overlapping periods could be used. However, we were mainly interested in finding out which recall period, starting from zero days recall, can be recommended to be used in the future when measuring productivity costs, given the idea that only one question should be used to measure absence from work retrospectively. Therefore, analysis of adjacent periods is not meaningful.

In our study we did not find a systematic difference between registered and reported sick leave. This, however, should not lead to the conclusion that systematic over- or underestimation does not exist in self-reported days absent due to illness. As mentioned before, the number of cases in our study might be too small to find a significant difference. Given the mean difference above zero, the reported sick leave seems to be less than the company registered absence. In case a systematic difference in fact might exist, this does not necessarily lead to the conclusion that recall bias exists when measuring sick leave

retrospectively. As long as imperfect memory (both by chance or systematic) is equal in the different groups under comparison in a clinical trial, recall bias is not an issue. However, respondents are less likely to recall the absence of interest when the disease under investigation is no longer present or treatment is successful. If this is the case, significant research findings based on retrospective data might be interpreted as a methodological artefact (Raphael, 1987). Measuring sick leave retrospectively as part of an economic evaluation comparing different treatment options might introduce recall bias when a difference in treatment success is found. Therefore, one should be aware using rather long recall periods.

To our knowledge the study by Burdorf *et al.* (1996) explicitly concerned reliability of a questionnaire on sickness absence. Concordance between company-registered and reported data were analysed by calculating Cohen's κ. Cohen's κ is based on the distinction between agreement between observations on the basis of chance, and the actual agreement between observations beyond chance (Landis & Koch, 1977). The number of days absent due to illness was categorised into four categories and, based on these figures, four by four tables were determined comparing registered and reported sick leave. However, we did not use this method because this approach has its' limitations. For instance, a perfect κ can be calculated where respondents for whom registered data indicate 50 days and the reported number of days is only 5 days as long as both numbers of days belong to the same category used in the analysis. Thus, when using κ it should be clear that after classifying the data in a limited number of categories, the absolute difference between registered and reported data is not taken into account as is done in our method.

Several limitations must be considered when interpreting the results of our study. First, it was not possible to investigate the non-response. In the study by Burdorf *et al.* (1996) it was found that the prevalence and duration of sick leave was higher among non-respondents than respondents. They conclude that the biased response might be due to the fact that, as in our study, all workers were informed of the purpose of the study. Although confidentiality was guaranteed, this may have caused workers with high levels of absence due to illness to refrain from participating. Our findings were not different between analyses regarding all respondents and cases. This seems to indicate that the non-response probably does not influence our conclusion. Second, the recall periods that were used in our study were limited to five periods. From our analyses it seems clear that a recall period of 6 months or more might lead to recall bias, and a recall period of 2 months or less might not lead to recall bias. However, based on our data it is not possible to judge recall periods in-between 2 and 6 months. Third, compared to Dutch sick leave figures our study group seems to have a low rate of absence due to illness (Statistics Netherlands, 1997b). One of the reasons for this deviation might be the fact that the company that participated in our study existed largely of office workers. In the study by Burdorf *et al.* (1996), however, no significant difference was found between office workers and blue-collar workers regarding the reliability of retrospective measurement of sick leave.

In conclusion, in the view of accurate measurement of costs and consequences, a retrospective questionnaire used for measuring absence from work due to illness can be a

reliable source of data. However, researchers should be aware of imprecision related to the recall period used. Because precision of 6 and 12-month recall is very poor when measuring productivity costs retrospectively, we recommend using a recall period of no more than two months.

ACKNOWLEDGEMENTS

The authors are grateful to Mrs. I. Van Camp M.Sc. and Mr. H. Nelis for data collection, Mr. W. Lemmens, and Mrs. L. Lemmens for data management, and Mrs. P. Pasker PhD. for comments. This study was supported by a grant from Astra Pharmaceutica BV, The Netherlands

CHAPTER 5
THE ISSUE OF CREDIBLE VALUING OF COSTS AND CONSEQUENCES

CHAPTER 5.1
THE ISSUE OF CREDIBLE VALUING OF COSTS AND CONSEQUENCES: TAKING SELF-REPORTED COMPENSATING MECHANISMS INTO ACCOUNT WHEN CALCULATING PRODUCTIVITY COSTS

Based on Severens JL, Laheij RJF, Jansen JBMJ, Lisdonk EH van de, & Verbeek ALM (1998). Estimating the cost of lost productivity in dyspepsia. *Alimentary Pharmacology and Therapeutics* 12: 919-923.

INTRODUCTION

Credible valuing is a topic which is discussed regularly when including productivity costs in economic evaluations of medical technologies (Luce *et al.*, 1996; Brouwer *et al.*, 1997a; Weinstein *et al.*, 1997; Brouwer *et al.*, 1997b). This disagreement is illustrated clearly by the fact that the Canadian guidelines for economic evaluation (Canadian Coordinating Office for Health Technology Assessment, 1997) suggest inclusion of these costs, in contrast to the revision of the Australian guidelines (Langley, 1996) which suggest to exclude these costs. One of the major concerns for not including productivity costs is the way these costs are predominantly valued. The currently used approach for calculating productivity costs due to absence from work is to take days absent valued by gross earnings, using the argument that this reflects the lost value of production when individuals are absent from work (Koopmanschap & Rutten, 1996b). This in fact reflects the potential production costs, while the actual costs for society may be lower because the absence is compensated for (Koopmanschap & Rutten, 1996a). For the long term absence, work can be taken over by an unemployed person or by reallocating employees over jobs. In contrast to the traditional human capital approach[8], the so-called friction cost method takes these kind of compensating mechanisms into account (Koopmanschap *et al.*, 1995). For short-term absence, a person's work may be covered by others, postponed, or made up by the sick person on their return to work. These compensating mechanisms influence the estimation of the productivity costs.

From previous studies concerning various topics in gastroenterology it can be concluded that a societal perspective is obviously a relevant choice when performing an economic evaluation (Jönsson & Karlsson, 1996). This perspective implies incorporating all relevant costs for patient management into the study and thus, in gastrointestinal disease the productivity costs appear to be highly relevant. This is confirmed by the fact that work loss is substantial (up to a mean of 24 days in 1 year) among employed dyspeptic patients (Jensen, 1988; Johannessen *et al.*, 1990; Jönsson & Carlsson, 1991; Crean *et al.*, 1994). Using days absent from work to calculate productivity costs, it was found that in the case of dyspepsia these costs were by far the most dominant (up to 79% of the total costs) compared to other types of costs (Pym *et al.*, 1990; Nyrén *et al.*, 1985; Sonnenberg & Everhart, 1997). In one study, a net economic gain was calculated for treating dyspeptic patients because of the reduction in productivity costs. In this case extra costs of medical treatment were outweighed by savings in productivity costs (Bytzer *et al.*, 1994). However, the question arises as to whether these costs may be overestimated because compensating mechanisms for absence from work were not taken into account.

8. The human capital approach values lost productivity by estimating the discounted future earnings of an individual in the years that they would have worked would they not have become ill.

In this study we analysed the productivity costs as a result of short-term absence from work due to dyspepsia, taking self-reported compensating mechanisms for absence into account. A comparison is made between the current method for calculating productivity costs and our approach, which limits overestimating productivity costs.

MATERIALS AND METHODS

Data

Two studies on the relative efficacy of different combined diagnostic and treatment strategies in patients with dyspepsia were carried out, using a questionnaire about absence from work due to dyspeptic complaints. Patients who visited their general practitioner for the first time because of dyspeptic complaints were enrolled in the first study (Study 1). Patients who visited their general practitioner and were known to have persistent dyspeptic complaints serious enough to warrant further diagnosis by gastrointestinal endoscopy were enrolled in a second study (Study 2). A detailed description of the inclusion criteria of Study 2 is reported elsewhere (Laheij *et al.*, 1998). In both studies, before treatment was started, patients filled in a questionnaire. Besides general information such as age and gender, respondents were asked to indicate the severity of their dyspeptic complaints. Level of education, profession, employment status, number of hours worked per week, and type of work (e.g. shiftwork, executive responsibilities) were also indicated by the respondents. Respondents were asked to report the number of days of missed work during the past 4 weeks due to either illness in general or to specific dyspeptic complaints. They were asked to indicate any compensating mechanisms in working activities for short-term absence. A choice could be made from six categories of compensating mechanisms: 1) compensation by colleagues during normal working hours; 2) compensation by colleagues during extra working hours; 3) compensation by extra temporary workers; 4) self-compensation during normal working hours; 5) self-compensation during extra working hours; 6) no compensation for lost work and compensating mechanisms not known.

Valuation of absence from work

The analysis of productivity costs concentrated on people who had a paid job during the last 4 weeks prior to the study. When calculating the productivity costs of a disease or a medical intervention due to absence from work, the number of days registered as being 'lost' must be multiplied by the productivity costs per day. The productivity costs are estimated by using the average gross wage per day, including employers' social benefit premiums for the Netherlands (i.e. gross earnings before deductions, plus employer-paid social benefits) (Statistics

Netherlands, 1997b). The logic here is that this reflects the value of production (Drummond *et al.*, 1997). The number of days absent from work due to illness in general and dyspeptic complaints in particular were multiplied by the average gross wage per day to estimate the productivity costs without taking compensating mechanisms into account. In our analysis, we did take compensating mechanisms into account. Two types of calculations were used. First, we used a conservative approach in which productivity costs were considered to be relevant only in cases where respondents answered that compensating mechanisms were used which needed extra financial efforts, such as overtime work by colleagues or using extra workers. Second, we used an adjusted approach in which productivity costs were calculated in the same way as in the previous approach except when a respondent answered that no compensating mechanism existed at all (work was 'lost') or work could be made up by the sick employee on his return to work. For these latter cases productivity costs were considered to be zero. In all other cases productivity costs were reflected by the employee's productive value, i.e. the average gross wage. In both approaches the value of the unknown financial effort was estimated by again using the average gross wage.

RESULTS

Table 5.1 Baseline characteristics of the respondents.

	study 1	study 2
males/females	22/31	42/39
mean and range of age (years)	49 (24 - 66)	44 (20 - 74)
mean and range of severity of dyspeptic symptoms	6.15 (0 - 12)	6.0 (1 -12)
employment status N (%)		
employed	21 (41%)	45 (62%)
employed unpaid work/housekeeping	18 (35%)	16 (21%)
disabled	7 (14%)	3 (4%)
unemployed	1 (2%)	4 (5%)
retired	4 (8%)	6 (8%)
employment characteristics N (%)		
employed full time	15 (71%)	30 (69%)
irregular working days	3 (15%)	13 (29%)
shift work	2 (10%)	14 (31%)
management staff	6 (29%)	8 (18%)

In total 136 patients were included in the studies, 54 in Study 1 and 82 in Study 2. Baseline characteristics of the patients for the two studies are shown in Table 5.1. The baseline

characteristics were no different between the studies with respect to gender, age and severity of dyspeptic complaints (scale from 1 to 12) Of the 136 patients studied, 66 had a paid job at the time of answering the questionnaire, 34 performed housekeeping or did volunteer work, 14 were unable to work or were unemployed, 10 were retired, and one was a student (11 missing values)

Of the patients from both studies who had a paid job, 25 (38%) reported absence from work which resulted in an average of 3 0 days in the past 4 weeks Work absence of 1 9 days was reported as due specifically to dyspeptic complaints All working patients answered the question about compensating mechanisms when absent from work The results of compensating mechanisms for each study are reported in Table 5 2

Table 5 2 Compensating mechanisms reported by respondents with a paid job

	study 1 (N=21)	study 2 (N=45)
compensation by colleagues during normal working hours	12 (57%)	21 (50%)
compensation by colleagues during extra working hours	0 (0%)	2 (5%)
compensation by employing temporary workers	6 (28%)	8 (19%)
self-compensation during normal working hours	1 (5%)	3 (7%)
self-compensation during extra working hours	1 (5%)	2 (5%)
no compensation for lost work	1 (5%)	6 (14%)
compensating mechanisms not known	0 (0%)	0 (0%)

On the basis of the days reported absent from work the productivity costs were calculated using the average gross wage for the Netherlands (Table 5 3) Our new conservative approach gave a much lower estimate of the productivity costs both in general and for productivity costs related to absence from work because of dyspeptic complaints Only one-quarter of the cost level calculated using the current approach remained, taking into account compensating mechanisms for absence from paid work The calculated range remained the same for the productivity costs Using the adjusted approach for taking the reported compensating mechanisms into account, only a marginal difference was found with the current approach of valuing each day of absence as productivity costs

Discussion

Based on the assumption that production falls because of absence from work, work loss has been reported to be substantial among employed dyspeptic patients (Johannessen *et al* , 1990, Crean *et al* , 1994, Sonnenberg & Everhart, 1997) A database on dyspepsia from the Southern General Hospital in Glasgow shows that each employed patient with functional dyspepsia

loses on average 18.3 weeks of work per year (Crean *et al.*, 1994). In a Swedish study it was found that patients with functional dyspepsia were responsible for, on average, 26 more days of lost production per year than other employees (Nyrén *et al.*, 1985). Compared to these findings our results show that absence from work due to dyspepsia in our population is relatively low, calculated on the basis of the reported days absent from work during the previous 4 weeks. During the measurement of absenteeism data a definite diagnosis of the cause of the dyspeptic complaints had not yet been made. It is expected that the actual cause does not influence absenteeism (Johannessen *et al.*, 1990). A comparison with other studies examining costs in gastroenterology should be carried with caution, because the aim of our study was only to compare different methods for calculating productivity costs.

Table 5.3 Days absent from work and productivity costs due to short-term absence during a 4 week period; mean and (range).

	sickness in general		dyspeptic complaints	
	study 1 (N=21)	study 2 (N=45)	study 1 (N=21)	study 2 (N=45)
days absent from work during past 4 weeks	3.1 (0 - 20)	3.0 (0 – 20)	1.2 (0 - 8)	2.2 (0 - 20)
productivity costs (in Dutch guilders)[9]				
not taking compensating mechanisms into account (current approach)	900 (0 - 5.899)	898 (0 - 8.259)	354 (0 - 2.360)	636 (0 - 8.259)
taking compensating mechanisms into account:				
conservative approach	197 (0 - 4.129)	256 (0 - 8.259)	56 (0 - 1.180)	249 (0 - 8.259)
adjusted approach	900 (0 - 5.899)	865 (0 - 8.259)	354 (0 - 2.360)	603 (0 - 8.259)

Besides absenteeism from work, productivity costs can be caused by reduced productivity of employees who work while suffering from symptoms of a specific disease. Until now two methods have been used to estimate productivity costs while working (Osterhaus *et al.*, 1992; van Roijen *et al.*, 1996). There is considerable uncertainty about the preferred valuation method because results vary considerably between different methods. Because of this we did not consider measuring productivity costs of employees working with dyspeptic complaints. In addition, the present study does not consider costs of absenteeism other than productivity

9. The 1997 mean exchange rate of one Dutch Guilder is approximately 0.31 British pounds and 0.51 US dollar.

costs. Such costs, for example administrative costs, undoubtedly exist (Koopmanschap & Rutten, 1996a).

Recommendations

Because productivity costs incurred or avoided when treating dyspeptic patients are substantial, this type of cost should be incorporated in economic evaluations of gastroenterologic interventions. When absenteeism from work is measured in work days lost and these days are translated into productivity costs, overestimation can be avoided. A general approach to take compensating mechanisms into account is part of the friction cost method. Numerous studies have shown that a reduction of annual labour time causes a less than proportional decrease in labour productivity (estimates between 0.6 and 0.9). To take short term absence compensating mechanisms into account, Koopmanschap *et al.* (1995) assume the elasticity to be 0.8 during the friction period, which is the period needed to replace a sick employee. Compared to our findings this estimated elasticity seems to be high. We calculated productivity costs as a result of absence from work due to illness in general and due to dyspeptic complaints, taking reported compensating mechanisms into account. The exact relationship between absence from work and real production loss depends on the employee's profession, the type of organisation and the production process (Koopmanschap & Rutten, 1993). This suggests that a detailed study at the company level would be necessary, but when analysing the costs involved with a specific disease or medical intervention this appears to be impossible. Therefore, patients' questionnaires must be used to estimate productivity cost. In our study all patients who had a paid job were able to indicate the number of days absent from work and to respond to the question if compensation mechanisms were used during their absence, and if so, of what type (no missing values). This indicates that it is possible to gather information about postponing work and the extent of internal labour reserves and flexible labour for each patient's specific situation. We can conclude that in taking reported compensating mechanisms into account when calculating productivity costs, estimates are significantly lower compared to the current method. It is not possible to judge the validity of the different methods because true productivity costs can only be measured within the specific employment situation of each respondent. Further research on this topic should be developed before it can be decided which method is preferable.

In conclusion, with respect to the principle that productivity costs should reflect actual production loss, in our opinion merely the absence from work leads to an overestimation and therefore compensating mechanisms should be taken into account. Complete insight of the consequences of compensating mechanisms requires detailed studies on the level of the firm which seem to be impossible when evaluating medical technologies. An approach to handling uncertainty in cost analysis related to health care interventions is to use sensitivity analysis (Briggs *et al.*, 1994). With sensitivity analysis it is possible to explore the implications of selecting a particular analytic method from several alternatives. Analogous to the principles of

sensitivity analysis, both the current and conservative approach should be used when estimating productivity costs. This gives insight into the impact of the different calculating methods for calculating productivity costs on the final conclusions of the study. In this way, overestimating productivity costs can be avoided.

ACKNOWLEDGEMENTS

The authors thank Mr. L. Van Rossum, Department of Epidemiology and Mr. J. Mulder Department of Medical Statistics, University of Nijmegen for data-management. Part of this project was financially supported by Astra Pharmaceutica BV, The Netherlands.

CHAPTER 5.2
THE ISSUE OF CREDIBLE VALUING OF COSTS AND CONSEQUENCES: WILLINGNESS TO PAY FOR NON-DECISIONAL DIAGNOSTIC INFORMATION

Based on Severens JL, Boo ThM de, Roosmalen MS van, Verweij PE & Wilt GJ van der. Validity of willingsness-to-pay for non-decisional diagnostic information [submitted for publication].

INTRODUCTION

Cost benefit analysis is a type of economic evaluation in which the consequences or outcome of a facility is expressed in monetary units. This makes it possible to compare costs and benefits in a direct way and to calculate the net social benefits. Thus, the goal of cost-benefit analysis is to identify whether a programme's benefits exceed its costs, indicating that a programme is worthwhile (Drummond *et al.*, 1997). Although several methods exist for estimating money values for health care programmes, the willingness-to-pay (WTP) survey techniques, known as contingent valuation, are popular (O'Brien & Gafni, 1996). A literature review regarding contingent valuation studies shows that the number of such studies is growing rapidly and that the majority are done as part of cost-benefit analysis (Diener *et al.*, 1998). Considering the economic evaluation of diagnostic facilities, however, not many studies were based on the principles of cost-benefit analysis (Severens & van der Wilt, 1999b).

Diagnostic technologies and procedures in general are typically judged effective if they provide information which is relevant related to a treatment/no-treatment decision and eventual patient outcome (Fryback & Thornbury, 1991; Severens *et al.*, 1999a). Nevertheless, information which is not directed towards therapeutic decision making might have a value to either physician or patient. This value can be defined as the non-decisional value of diagnostic information (Woodward *et al.*, 1998). In the literature several methods to measure this value can be found varying from open-ended questions asking how respondents felt about knowing their diagnostic results, to health status measurement using questionnaires such as those on anxiety and the SF36, to WTP (Mushlin *et al.*, 1994; Ried, 1994; Woodward *et al.*, 1998). How to measure the non-decisional value of diagnostic information remains uncertain. In this study we addressed the question whether WTP analysis is a method which leads to credible valuing of non-decisional diagnostic information.

An approach to investigate the validity of using a questionnaire when there is no criterion standard is to examine construct validity. A construct is a theoretically derived notion of the domain the instrument should measure, leading to expectations about how an instrument should behave if it is valid (Spilker, 1996). To examine the construct validity of the WTP procedure as a method to assess the value respondents place on a non-decisional diagnostic information, we tested the following hypotheses:

1. Subjective importance of testing, reflecting the risk-aversion of a person, is positively related to the WTP for diagnostic information (Woodward *et al.*, 1998);
2. The burden of testing experienced is negatively related to the WTP for diagnostic information (Appel *et al.*, 1990);
3. Subjective belief in the accuracy of a diagnostic test is positively related to the WTP for diagnostic information;

4. Prior knowledge of the presence or absence of the disease is negatively related to the WTP for diagnostic information (Asch *et al.*, 1990);
5. The severity of the disease tested (perceived risk) is positively related to the WTP for diagnostic information (McDaniels *et al.*, 1992; Lindholm *et al.*, 1997; Kobelt, 1997); and
6. The possibility of treatment after testing is positively related to the WTP for diagnostic information (Donaldson *et al.*, 1995).

Histoplasmosis

The hypotheses were tested among individuals who are considered to be at increased risk of histoplasmosis. Histoplasmosis is caused by the dimorphic fungus *Histoplasma capsulatum*. The organism grows in its mycelial phase in soil, especially in soil that has been nitrogen-enriched by bat or bird guano. Infection is acquired when microconidia are inhaled into the lungs where they transform into the pathogenic yeast-phase organisms (Houston, 1994; Gurney & Conces, 1996). The organism is widely distributed throughout the world in both tropical and temperate climates. However, histoplasmosis is considered a rare disease in Europe (Houston, 1994) and the vast majority of European cases are attributable to endogenous reactivation of a latent infection acquired in overseas endemic areas (Manfredi *et al.*, 1994; van Crevel *et al.*, 1997). Several outbreaks of histoplasmosis have been reported to be cave-associated and therefore speleologists may be at increased risk of acquiring cave-associated histoplasmosis (Sacks *et al.*, 1986; Johnson *et al.*, 1988; Noel *et al.*, 1995; Suzaki *et al.*, 1995).

METHODS

Respondents

A group of Dutch speleologists was invited to participate in the study during their annual membership meeting. Each participant was asked to complete a self-administered questionnaire and blood samples were obtained. The questionnaire focused on the history of spelunking activities to identify the possible risk factors of acquiring cave-associated histoplasmosis. Furthermore, the participants were asked about clinical symptoms suggestive of histoplasmosis (e.g. a non-productive cough, substernal pain and shortness of breath) after spelunking in a particular area and about having had a previous test for histoplasmosis. The blood samples were tested for the presence of antibodies to *H. capsulatum*. A detailed description of the study is presented elsewhere (van Roosmalen *et al.*, 1998).

Questionnaire

Because treatment for histoplasmosis is not indicated in an asymptomatic person, for the speleologists the blood sample would give non-decisional diagnostic information. Currently, testing for the seroprevalence of antibodies to *H. capsulatum* is not part of regular health care in the Netherlands. Testing among a risk population and measuring their WTP can be described as an ex post user based perspective (O'Brien & Gafni, 1996). This perspective implies asking individuals who would potentially gain from using the specific medical technology what maximum amount in monetary units they would pay to gain access to the facility. Different methods for asking somebody's WTP can be found in the literature. We chose to frame the WTP question using five answering categories. We did so because simply asking the maximum a respondent would be willing to pay poses a large cognitive task for a respondent (O'Brien & Gafni, 1996). Besides this, when open-ended questions are used, an important concern for WTP analysis is the treatment of outliers (Berwick & Weinstein, 1985; McDaniels *et al.*, 1992) and open-ended questions are likely to be biased and erratic (Donaldson *et al.*, 1995). Furthermore, because income is an important predictor variable in WTP analysis (McDaniels *et al.*, 1992; Johannesson *et al.*, 1997), we tried to eliminate the possible influence of this variable by defining the WTP question using a percentage of the respondents' monthly income. Regarding these considerations, we measured WTP using five answering categories ranging from not willing to pay anything to willing to pay an increasing maximum percentage of the respondents' monthly income (1%, 5%, 10%, and more than 10%) were defined in the WTP question.

Table 5.4 Re-coding of the original variables into dichotomous variables.

original five category variable	dichotomous variable
willingness to pay	not willing to pay anything - willing to pay some amount
subjective importance of testing	not very important - very important
the burden of testing	no high burden of testing - high burden of testing
perceived reliability of the test	not very reliable - very reliable
possibility of being infected	no strong idea about infection - strong idea about infection
perceived severity of the disease	not a severe disease - a severe disease
perceived possibility to treat the disease	not possible to treat - possible to treat

In order to test the hypotheses, respondents were asked to indicate 1) the subjective importance of testing, 2) the burden of testing (taking a blood sample), 3) the perceived reliability of the test, 4) the subjective believe of being infected, 5) the perceived severity of the disease, and 6) the perceived possibility to treat the disease (from this point called the hypothesis-related variables). For each question a five category scale was used. For the

purpose of performing further statistical analyses (avoiding the problem of near empty cells and thus also introducing too many dummy variables in logistic regression analysis) both the WTP variable and the hypothesis-related variables were recoded into dichotomous variables (Table 5.4).

Analysis of the data concerned several aspects. All participants were generally described for age, gender, and years active in spelunking. Frequencies of the answering categories to the WTP and hypothesis-related questions were calculated. Spearman correlation coefficients were calculated between the WTP variable and the six hypothesis-related variables. To examine univariate relationships we used the chi-squared test on the re-coded variables. Multivariate dependence of WTP with the recoded variables was investigated using stepwise logistic regression. The level of significance used was 5% .

RESULTS

Of the 90 attendees at the membership meeting, 84 agreed to participate. Of the 84 participants, 67 were male (80%) with a mean age of 38.6 years (range 20-62), and 17 (20%) were female with a mean age of 33.5 years (range 23-45). The mean number of years active in spelunking was 13 (range 2-32). Of the 84 respondents, eight did not answer the question on WTP for the *H. capsulatum* test and therefore the number of respondents for our analysis was 76.

Table 5.5 Number and percentage of respondents regarding their willingness to pay for non-decisional diagnostic information (N = 76).

answering category	number (percentage)
not willing to pay anything	22 (29%)
willing to pay maximum of 1% of monthly income	39 (51%)
willing to pay maximum of 5% of monthly income	12 (16%)
willing to pay maximum of 10% of monthly income	3 (4%)
willing to pay more then 10% of monthly income	0 (0%)

Table 5.5 gives an overview of the number of respondents for each of the answering categories to measure the WTP for the diagnostic test. Twenty-two respondents (29%, 95% confidence interval (CI): 18.7% – 39.2%) indicated not to be willing to pay anything for the diagnostic information, as 54 respondents (71%, 95% CI: 60.9% - 81.2%) indicated to be willing to pay some amount for the diagnostic information. Testing our hypotheses resulted in a significant relationship between respondents' WTP and the subjective importance of testing. No univariate relationships were found between the WTP variable and the other five hypothesis-related variables (Table 5.6). Using the stepwise logistic regression to investigate multivariate dependence of WTP with the recoded variables, this resulted in a logistic model

only with the hypothesis-related variable subjective importance of testing. No other hypothesis-related variable, or any other variable such as age and gender of the respondent entered the model using the 5% significance level.

DISCUSSION

The purpose of our study was to examine the issue of credible valuing of non-decisional diagnostic information by measuring WTP. This issue was studied in asymptomatic healthy volunteers at increased risk for histoplasmosis. We were able to find a relationship between the subjective importance of testing and the respondents' WTP; the remaining five hypotheses, however, could not be confirmed. This suggests that WTP measurement may not be a valid method to assess the value that respondents place on non-decisional diagnostic information.

Table 5 6 Percentage of respondents indicating to be willing or not willing to pay for diagnostic information regaring the hypothesis related questions and the results of the chi-squared test.

hypothesis-related variable	willing to pay		p-value
	no	yes	
subjective importance of test. very important	23	52	0 02
burden of testing: high burden of testing#	52	48	0 74
Reliability of the test: very reliable	50	57	0 56
possibility of being infected strong idea about infection	86	85	0.90
severity of the disease. a severe disease	59	43	0.19
possibility to treat the disease· possible to treat#	27	26	0 94

Because of missing values N=75 instead of N=76.

In an early study by Thompson (1986) it is stated that a high response rate indicates overall feasibility of the method to measure WTP. Given a response rate of more than 90%, our study population seems not to have had any difficulty answering their WTP for diagnostic information. Golan & Shechter (1993) measured WTP for changes in the Israeli health care system. Because the WTP estimates were 'reasonable', they conclude that the method seems to adhere to the conditions for reliability and validity. Chestnut *et al.* (1996) measured the WTP for changes in symptoms in patients suffering angina. Actual expenditures and perceived angina episodes avoided and WTP for a hypothetical treatment led to comparable results and therefore the authors concluded that the WTP method was valid. Flowers *et al.* (1997) showed that test-retest reliability was high, more than 80% of the healthy volunteers participating in the study reported the method to be reasonable for therapeutic decision making, and 62% expressed comfort in using the method for their own health care decisions.

From these the authors concluded that the WTP method is valid. In a recent study by O'Brien *et al.* (1998) Health Maintenance Organisation members were asked their WTP for a new drug, using a user based scenario (assuming they were about the moment to use the drug) and an insurance based scenario (at risk for the disease in the future). Construct validity was concluded to be encouraging.

Our results do not mesh with the results from other studies as described above. A possible explanation for our findings is the fact that these studies were performed in several countries (United States, Canada, Israel) which do not have a health care financing system comparable to the system in the Netherlands. In the Netherlands the system of a comprehensive health insurance is the basis for financing health care facilities. Except for (luxury) health care interventions which are not covered by the basic insurance package, patients are hardly ever confronted with paying health care facilities out of pocket. Therefore, asking a persons' WTP for something one is not used to paying for, might be defined as an unrealistic hypothetical scenario leading to invalid results (O'Brien & Gafni, 1996). In the literature only one study using WTP measurement was found that was performed in the Netherlands. Krabbe *et al.* (1997) asked healthy student volunteers to imagine that they were in a certain impaired health state and asked their WTP to return to their previously healthy condition. The authors conclude that the reliability of WTP measurement was low and the number of inconsistencies was substantial. Although their method for asking the respondents WTP is different from the method we used, these results agree with our findings regarding the construct validity of the WTP measurement. It seems that WTP measurement as a means to value non-decisional information of diagnostic tests is not a valid method, especially when used with respondents who are not used to actually paying for health care services.

Our study has several limitations. First, the number of respondents is limited. Although a significant relation between the WTP and the subjective importance of testing was found, for the other variables to test our hypothesis, this relationship could was not found. The question arises if the power of our study based on these numbers is sufficient. Indeed, with sample sizes of 22 and 54, the difference in percentage 'success' between the two groups has to be at least 30% in order to be detectable with a power of 80%. A second aspect which might have influenced our results is the selection of respondents. Although an ex post user based perspective was used and therefore only individuals who were at increased risk were included in our study, asking the same WTP and hypothesis related questions at individuals not at risk for histoplasmosis might have led to more contrast in our findings. In the literature it is described that asking persons at risk (potential users of the facility) leads to higher WTP results than asking non-users (O'Brien & Gafni, 1996). However, in our case of testing for the seroprevalence of antibodies to *H. capsulatum* among a risk population it is practically impossible to identify a population of potential losers when this diagnostic facility in reality would be introduced in health care. Third, different methods to ask for somebody's WTP for non-decisional diagnostic information might lead to different results. Overestimating respondents' WTP might be the case because the question is hypothetical while actual

willingness to pay can be expected to lead to more moderate results (O'Brien & Gafni, 1996). Besides this, using answering categories instead of open-ended questions might also lead to higher WTP results (Donaldson *et al.*, 1995). However, because our study does not give any clear predictive factors for the WTP of the respondents, our findings give no indications for an overestimation of the WTP of our respondents.

Diagnostic information might be of value, even if it does not affect treatment decisions (Woodward *et al.*, 1998). In such cases however, the question remains how such non-decisional diagnostic information should be valued. In conclusion, using the concept of construct validity, we examined whether WTP measurement is a suitable method. Given the results, we were not able to conclude that WTP leads to credible valuing of non-decisional diagnostic information. We suggest that this may be primary related to the fact that as Dutch citizens, the respondents are not familiar with paying health care facilities out of pocket. This would mean that WTP is not the method of choice when assessing the monetary value that subjects place on health care interventions in countries with comprehensive coverage schemes. Further development of this method is clearly warranted.

ACKNOWLEDGEMENTS

The authors would like to thank the members of Speleo Nederland who participated in the study. Dr. B. O'Brien from McMaster University, Hamilton, Canada is thanked for helpful suggestions. This study was made possible by a grant from the Reinier Post Foundation.

CHAPTER 6
THE ISSUE OF UNCERTAINTY OF THE RESULTS: STATISTICAL ANALYSIS OF INCREMENTAL COST-EFFECTIVENESS RATIOS

Based on Severens JL, Boo ThM de & Konst EM (1999). Uncertainty of incremental cost-effectiveness ratios: a comparison of Fieller and bootstrap confidence intervals. *International Journal of Technology Assessment in Health Care* 15: 608-614.

INTRODUCTION

Increasingly, cost-effectiveness analyses are conducted as part of clinical trials (Glick, 1995; Drummond & Davies, 1991). Gathering data on both effectiveness and cost for each patient makes it possible to determine the mean and standard deviation of the different variables. This makes the issue of uncertainty of the results of an economic evaluation relevant.

Although some debate exists about the usefulness of average cost-effectiveness ratios of each alternative, the incremental cost-effectiveness ratio comparing an often new treatment or diagnostic facility to some alternative is considered to be the main result of cost-effectiveness analysis (Johannesson, 1995a; Briggs & Fenn, 1997a; Laska *et al.*, 1997b). Because the true costs C and effects E of the medical interventions to be compared (respectively 1 and 0) are not known for the population, the true incremental cost-effectiveness ratio R can not be determined. Information about the sample costs and effects are used to calculate the sample incremental cost-effectiveness ratio. This incremental cost-effectiveness ratio is defined as:

$$\hat{R} = \frac{\overline{C}_1 - \overline{C}_0}{\overline{E}_1 - \overline{E}_0}$$

However, $\hat{R}$ is a point estimate and reporting confidence intervals of at least the differences in costs and effects in addition to this estimate is advisable (O'Brien *et al.*, 1994; van Hout *et al.*, 1994). This in itself does not give any real insight into uncertainty of $\hat{R}$ itself. Several methods have been explored to estimate a confidence interval for R. O'Brien *et al.* (1994), van Hout *et al.* (1994), and Wakker & Klaassen (1995) use parametric methods for estimating confidence intervals for R. Willan & O'Brien (1996) concentrate on Fieller's theorem and Laska *et al.* (1997a) compare the Fieller method with two-sided Bonferoni confidence intervals. Briggs *et al.* (1997b) describe four methods for determining confidence intervals for R based on nonparametric bootstraps using different numbers of bootstrap replicates. Polsky *et al.* (1997) and Chaudhary & Stearns (1996) have compared both parametric and nonparametric methods for determining a confidence interval for R. Polsky *et al.* (1997) recommend the percentile method based on nonparametric bootstrap replicates and the parametric Fieller method. Robustness to distributional variation in costs and effects and the correlation between them are taken into account in these methods. Chaudhary & Stearns (1996) conclude that the key consideration in choosing between methods should be the extent to which the data are consistent with the assumptions behind the methods. A recent publication by Tambour & Zethraeus (1998a) compares the percentile bootstrap method with the method introduced by Wakker & Klaassen (1995). The bootstrap method is preferred.

The above mentioned articles appear to agree that methods which fully rely on a normal distribution of the incremental cost-effectiveness ratio should be avoided in determining confidence intervals of R. Different methods of calculating nonparametric

bootstrap intervals and the Fieller method fulfil this criterion. In this article we report our experiences comparing Fieller intervals and three methods for calculating bootstrap intervals: the percentile method and two Bias Corrected and Accelerated (BCA) methods.

METHODS

Trial data and simulation

The data used for comparing he methods in estimating confidence intervals was gathered in the framework of the Dutch PSOT study. The PSOT study was designed as a two-group randomised multicentre clinical trial concerning orthodontic treatment of children born with unilateral cleft lip and palate. One group of patients received presurgical orthopaedic treatment (PSOT, N=10), and the other group did not (non-PSOT, N=10). All other interventions were the same. Costs reported in this article are solely direct medical costs related to the PSOT intervention during the children's first year of life. A professional judgement of the overall speech and language performance (10 point scale) at the age of 2.5 year was used as a measure for treatment effectiveness. More detailed information about the PSOT study is reported elsewhere (Severens *et al.*, 1998).

From the trial data, for each group mean and variance of cost and effectiveness, and the correlation between them were computed. These statistics were used to simulate a subsequent 10,000 trials by repeatedly drawing 2 times 10 cases from normal distributions with these statistics as parameters. Results were rounded to the same number of decimal places as in the original trial. For each simulation, $\hat{R}_s^*$ was calculated; when the difference in effect equalled zero, $\hat{R}_s^*$ was made missing (unknown). The mean ICER $\overline{R}_s$ was calculated, and this value was used as estimate for the true population ICER, R (Figure 6.1).

Confidence intervals of the ICER

Fieller theorem method

The Fieller theorem method is a parametric method for calculating a confidence interval of a ratio of means. The assumption on which this method is based is bivariate normality of numerator and denominator, here the difference between the means of the cost and the means of effect (Laska *et al.*, 1997a). Let ΔE and ΔC denote the mean difference in effect and cost respectively, $S^2_{\Delta E}$ and $S^2_{\Delta C}$ their estimated variances and r the estimated (Pearson) correlation coefficient between them. Let $f_{v,1-\alpha}$ denote the upper percentage point of the F-distribution, with 1 and v degrees of freedom, v being the number of degrees of freedom upon which the estimated variance of $\Delta E - \Delta C$ is based (18 in our case).

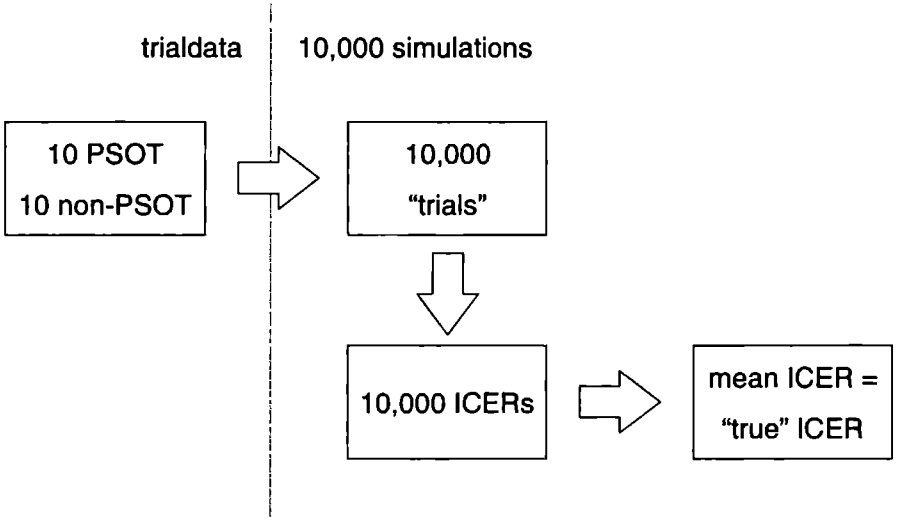


Figure 6.1 Overview of the relation between the trial data and the simulation data. The statistics of the original trial data were used to simulate a subsequent 10,000 trials. For each of the simulations the incremental cost-effectiveness ratio (ICER) was calculated. The mean ICER was used as estimate for the true population ICER.

Compute (Laska *et al.*, 1997a):

$$L_1 = \frac{(\Delta E\Delta C - f_{v,1-\alpha} r s_{\Delta E} s_{\Delta C}) - [(\Delta E\Delta C - f_{v,1-\alpha} r s_{\Delta E} s_{\Delta C})^2 - (\Delta E^2 - f_{v,1-\alpha} s_{\Delta E}^2)(\Delta C^2 - f_{v,1-\alpha} s_{\Delta C}^2)]^{1/2}}{\Delta E^2 - f_{v,1-\alpha} s_{\Delta E}^2}$$

and

$$L_2 = \frac{(\Delta E\Delta C - f_{v,1-\alpha} r s_{\Delta E} s_{\Delta C}) + [(\Delta E\Delta C - f_{v,1-\alpha} r s_{\Delta E} s_{\Delta C})^2 - (\Delta E^2 - f_{v,1-\alpha} s_{\Delta E}^2)(\Delta C^2 - f_{v,1-\alpha} s_{\Delta C}^2)]^{1/2}}{\Delta E^2 - f_{v,1-\alpha} s_{\Delta E}^2}$$

Now, if there is a statistically significant difference in effect, then (and only then) the denominators of L_1 and L_2 are positive. The Fieller $(1-\alpha)$ interval is then the interval (L_1, L_2). If there is no significant difference in effect, the denominators are negative and Fieller's interval consists of the union of the intervals $(-\infty, L_2)$ and $(L_1, +\infty)$. For further details see Laska *et al.* (1997a). (Note that in that paper there is a misprint on page 235: the second plus-sign in the numerator of the formula for the upper confidence limit estimator should be a minus-sign, as in the lower confidence limit). On the basis of each $\hat{R}_s^*$ of the 10,000 simulations, a 90% Fieller confidence interval was calculated.

Bootstrapping

The principle of bootstrapping is that a random sample of size n with replacement from the data is taken a large number of times (Efron *et al.*, 1993). Considering the fact that each $\hat{R}$ is estimated based on two replicated samples (treatment group and no treatment group) each simulation sample is used as a basis for the bootstrap replication which leads to, respectively, C_1^* and E_1^* for treatment, and C_0^* and E_0^* for control. As a result from each bootstrap series the bootstrap ratio $\hat{R}_b^*$ can be calculated:

$$\hat{R}_b^* = \frac{\overline{C}_1^* - \overline{C}_0^*}{\overline{E}_1^* - \overline{E}_0^*}$$

For each trial simulation we performed 25,000 bootstrap replicates to avoid problems when using a too small number of replicates. On the basis of the bootstraps replicates, confidence intervals for R can be calculated.

Bootstrap confidence intervals

We used three methods to calculate a bootstrap confidence interval. First we used the percentile method, which is based on the principle of sorting the $\hat{R}_b^*$. When 25,000 replicates have been made and the 90% confidence interval has to be determined, the percentile method uses the 1,250th and 23,750th ranked $\hat{R}_b^*$ as the confidence interval limits.

Another method for calculating confidence intervals based on bootstrap replicates is the so called bias corrected and accelerated (BCA) percentile method. The basic principle of the BCA method is a modification of the percentile method making a correction for bias and the skewness of the estimator of the sampling distribution (Briggs *et al.*, 1997b). For an extensive description of this method we suggest Efron & Thibshirani (1993). and Efron (1987). The BCA method uses an acceleration constant which is used to adjust for the skewness of the sampling distribution of $\hat{R}_b^*$. This acceleration constant is calculated using a jack-knife estimate. Since the jack-knife method is not straightforward described in case of comparison of two groups, we used two options: BCA-1 and BCA-2. For BCA-1 an estimator is used that leaves out only one measurement at a time, irrespective of the origin (treatment or no treatment). BCA-2 uses an estimator based on simultaneously leaving out one measurement from each of the two samples, treatment and no treatment.

Because in our study we simulated 10,000 trials, 25,000 bootstrap replicates were performed 10,000 times. Thus, for each of the three methods of calculating 90% confidence intervals based on the bootstrap replicates (percentile method, BCA-1, and BCA-2), 10,000 confidence intervals were determined (Figure 6.2).

Effect of $E_1 - E_0 \approx 0$ on the confidence intervals

Our original trial data showed a significant difference in effectiveness between the two groups of 1.34 on the scale of 1 to 10. To investigate the impact of the relative importance of the

magnitude of the difference in effectiveness between the groups, we simulated four other situations: the difference in effectiveness was assumed to be 0.5 and 1.0 less, and 1.0 and 5.0 more than in the original trial which, respectively, reflects the situations where $E_1 - E_0 \approx 0$ and where $E_1 - E_0 > 0$. Hence, in total five groups were used to compare the results of Fieller intervals and three bootstrap intervals; in all groups the same simulation data were used, except for subtraction from- or addition of the given constants to the (mean) difference in effect.

Adequacy of the confidence intervals

The adequacy of the different methods to calculate confidence intervals for R was investigated by comparing the confidence intervals containing $\overline{R}_s$, the mean of the 10,000 $\hat{R}_s^*$. The percentage of confidence intervals containing this estimate of the true R was determined for each of the four methods. Approximately 90% of the intervals should contain this estimate, because a level of miscoverage of 10% (type 1 error) was prespecified.

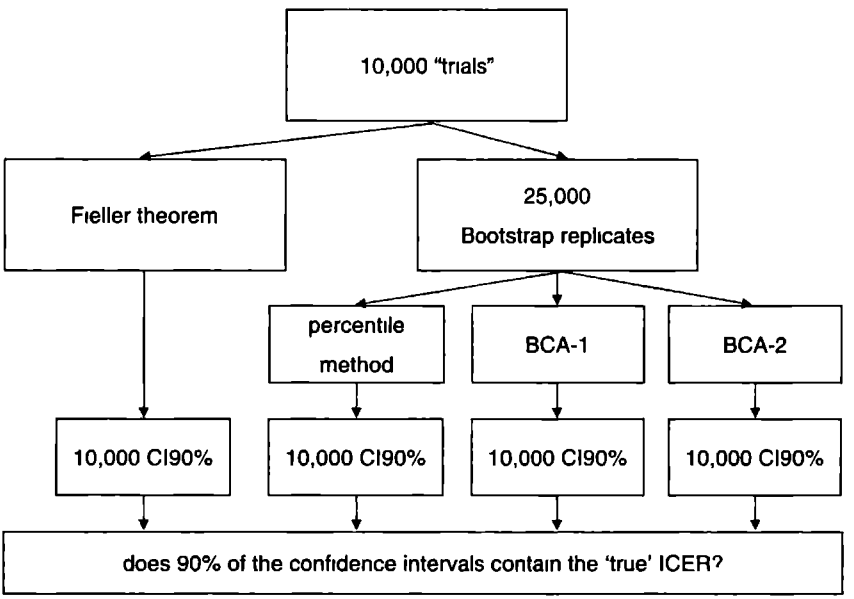


Figure 6 2 Overview of the simulation data, methods for calculating confidence intervals (Fieller theorem, Bootstrap percentile method, Bootstrap Bias Corrected and Accelerated 1 (BCA-1), and Bootstrap Bias Corrected and Accelerated 2 (BCA-2)) and calculation of the adequacy of these methods Based on the 10,000 simulated trials, for each simulation 90% confidence intervals were computed using the four methods The adequacy of these methods was investigated by calculating the percentage of confidence intervals containing the estimate of the true ICER.

RESULTS

Trial data

The trial concerned 20 patients (10 PSOT treated, 10 no PSOT treatment) For the treatment group the mean medical cost was Dfl 2,544 (Dutch guilders), standard deviation Dfl 646 and the mean effectiveness score (for speech and language development) and standard deviation were, respectively, 3 52 and 1 75 For this group, correlation between costs and effectiveness was 0 35 Mean costs and standard deviation for the no treatment group turned out to be Dfl 881 and Dfl 151 Mean effectiveness and standard deviation were respectively 2 18 and 0 62 Correlation between costs and effectiveness for no-treatment was -0 06

Combining the results of both groups, with the t-test the cost difference between the groups was significant Both the difference between the means of the costs (Dfl 1,663) and the means of effectiveness (1 34) were significant The correlation between difference in cost and difference in effects was + 0 30 The incremental cost-effectiveness ratio based on the trial data for the speech development score is Dfl 1,241 / point score improvement

Simulations

Table 6 1 Results of the simulations for five levels of difference in effectiveness

Level of difference in effectiveness	N	$\overline{R}_s$ (mean ICER)
baseline difference minus 1 0	9,935	2,388
baseline difference minus 0 5	9,982	2,692
baseline difference (1 34)	9,997	2,053
baseline difference plus 1 0	10,000	941
baseline difference plus 5 0	10,000	633

Table 6 1 summarises the result of the simulations for the chosen levels of difference in effectiveness Note that from the baseline difference of 1 34 in speech and language score the values 0 5 and 1 0 are subtracted to obtain the situation where $E_1 - E_0 \approx 0$ The baseline simulation and the baseline difference in effectiveness increased by 1 0 and 5 0, respectively, reflecting the situation that $E_1 - E_0 > 0$ From this table it can be seen that the number of simulated trials on which $\overline{R}_s$ is based is not always equal to 10,000 This is due to the fact that whenever a simulated trial leads to $E^*_{s1} - E^*_{s0} = 0$, $\hat{R}^*_s$ can not be calculated for this simulation and therefore is neglected when calculating $\overline{R}_s$ Whenever the difference in effectiveness is significant, the probability that one of the simulations leads to $E^*_{s1} - E^*_{s0} = 0$ gets (much) smaller and more simulations are used to calculate $\overline{R}_s$

Confidence intervals for $\hat{R}$

The results of determining the adequacy of the four different methods to compute confidence intervals for R are shown in Table 6 2 The percentages of the confidence intervals that contain $\overline{R}_s$, are mentioned for the four different methods, separately for the different levels of difference in effectiveness As can be seen from this table, in case $E_1 - E_0 \approx 0$ (baseline difference minus 1 0), both bootstrap based BCA-1 and BCA-2 calculations lead to confidence intervals that contain $\overline{R}_s$ in less than 70% of the cases In contrast, the bootstrap percentile method contains 96% of the time $\overline{R}_s$, indicating (much) too wide intervals This is not surprising in view of the fact that the denominator of R_b^* often is practically zero The Fieller method seems to have intermediate results, however, the results do not appear to be stable when comparing the situation regarding a baseline difference minus 0,5 (reflecting $E_1 - E_0 \approx 0$ to a lesser extent) and the baseline effectiveness difference (reflecting $E_1 - E_0 \neq 0$) In the situation when $E_1 - E_0 > 0$ (baseline difference plus 1 0 and 5 0 respectively), the target percentage of 90% coverage of $\overline{R}_s$ by the 10,000 confidence intervals is best obtained by the Fieller method All bootstrap based calculating methods give somewhat too narrow intervals, thus less than 90% contain $\overline{R}_s$

Table 6 2 Percentage of the 10,000 90% confidence intervals that contain the mean incremental cost-effectiveness ratio $\overline{R}_s$ for four methods of calculating the confidence intervals

level of difference in effectiveness	Fieller	bootstrap		
		percentile	BCA-1	BCA-2
baseline difference minus 1 0	75 9	96 5	69 2	60 6
baseline difference minus 0 5	89 0	90 9	77 0	72 8
baseline difference (1 34)	84 1	82 0	79 6	77 7
baseline difference plus 1 0	88 2	85 7	85 5	85 6
baseline difference plus 5 0	89 2	85 5	85 5	85 5

Discussion

We did not investigate the impact of varying the correlation between cost and effectiveness on our study findings Polsky *et al* (1997) state that the bootstrap percentile method seems to perform slightly better than the Fieller theorem method when correlation of cost and effectiveness was positive In case of a negative correlation this seems to be the other way around However, in this study based on bootstrap replicates only the percentile method was used for calculating confidence intervals Other parametric approaches were highly influenced by correlation between cost and effectiveness For comparison in our study, the correlation between difference in cost and difference in effects was + 0 30

Recently, an alternative approach was suggested which gives a solution to the difficulty in calculating confidence intervals for ratios (Tambour *et al.*, 1998b). In this approach the effectiveness units as used in a study are multiplied by the price per effectiveness unit, thus resulting in expression of both costs and effectiveness in monetary terms. Tambour *et al.* (1998b) describe that in this way the net benefits of the medical intervention which are compared can be determined and that standard statistical techniques can be used to calculate confidence intervals for the net benefits. However, we think that this approach in general is difficult to apply. First, the method makes it necessary to determine a price per effectiveness unit and this seems to be rather arbitrary. Although sensitivity analyse might be used to explore the impact of varying the unit price on the studies' conclusions (Briggs *et al.*, 1994) it can be argued that specific measures of effectiveness are difficult to express in monetary terms. Second, the example which is used by the authors expresses the effectiveness in QALYs which seems to be possible to translate into monetary terms quite easily. However, the question arises hoe to translate effectiveness other measures, because QALYs are only one of many alternatives to express effectiveness (Severens & van der Wilt, 1999).

In conclusion, to answer the question which method should be used to determine a confidence interval for an incremental cost-effectiveness ratio, - either the Fieller theorem method or one of the three bootstrap based calculations used in our analyses - a distinction must be made between the situation where $E_1 - E_0 \approx 0$ and where $E_1 - E_0 > 0$. Actually, in case a trial does not show a significant difference in effectiveness it does not make sense to calculate confidence intervals for R, making the discussion about which confidence interval method to use academic (Chaudhary & Stearns, 1996). In such cases it would be better to concentrate solely on cost differences. If in such situations nevertheless confidence limits are determined, as already noted by Briggs *et al.* (1997b), one should be very wary when trying to interpret bootstrap results. The same can be said of Fieller's limits, because such intervals consist of two parts: from minus infinity to a negative value and from a positive value to plus infinity. Such intervals are in practice of no use.

In the situation where $E_1 - E_0 \neq 0$, the different methods for calculating confidence intervals for R can be used but do not seem to give very different results. When using the bootstrap method it seems advisable to show the bootstrap results on the cost-effectiveness plane (Briggs *et al.*, 1997b). This clearly shows in which quadrant of the plane the 'population' ratio can be expected, which is important to decide about inferiority or dominance of one medical alternative to another. Calculating the confidence interval of this ratio is an additional step which should be applied only when the difference in mean effectiveness is large enough. However, bootstrap replication has the disadvantage that a rather powerful computer is necessary, while the Fieller theorem method, which is based on a single formula, can be applied without simulation techniques. And besides this, Fieller limits seem to approximate the target coverage of 90% better. It could be argued, that Fieller's limits appear to be the best in this study, because we sampled from normal distributions. We believe however, that in many practical situations both cost and effect differences can be very well

approximated with a normal distribution. Furthermore, even in situations where deviations from normality are non-negligible, if sample sizes are large enough (say, > 50) the central limit theorem guarantees a satisfactory approximation.

Recommendations

Based on our comparison of Fieller intervals and three methods for calculating bootstrap intervals - the percentile method and two Bias Corrected Accelerated (BCA) methods - the following recommendation can be given. First, a possible difference in effectiveness between groups compared should be tested with the t-test. When there is no significant difference in effectiveness it is absolutely not useful to calculate the incremental cost-effectiveness ratio. In such a case one should concentrate on estimating cost differences instead. Whenever there is a strictly significant difference in effectiveness between groups, calculating a confidence interval for the ICER is useful. Fieller's theorem leads to satisfactory results, and the necessary calculations are relatively easily done compared to bootstrap simulations. We therefore recommend using the Fieller confidence limits, in such cases.

CHAPTER 7
GENERAL DISCUSSION

In this thesis several methodological issues related to the economic evaluation of health care technologies were discussed: the choice of the competing alternative; the relevant costs and consequences; the accurate measurement of costs and consequences; credible valuing of costs and consequences; and the issue of uncertainty of the results of an economic evaluation. These and the other issues that are mentioned by Drummond *et al.* (1997) aim to assist users of economic evaluations in assessing the validity of the results they encounter. For each separate issue a specific aspect was studied. Besides the Drummond checklist, guidelines for economic evaluations exist in several countries. The purpose of these guidelines is to provide guidance on the appropriate concepts to consider or use when conducting such economic evaluations (Jacobs *et al.*, 1995). The aspects we studied are discussed below within the framework of guidelines for (pharmaco-) economic evaluation, including the guidelines from Canada (Canadian Coordinating Office for Health Technology Assessment, 1997), Ontario (Jacobs *et al.*, 1995), Australia (Access and Financing Division, 1998), United Kingdom (Towse, 1997), and The Netherlands (National Health Insurance Board, 1999).

The choice of the competing alternative: the sequence of diagnostic tests and modelling therapeutic alternatives

As is stated by Drummond *et al.* (1997), clear identification of the primary objective of the alternatives being compared is essential for the readers of the results of an economic evaluation to judge the applicability to their own settings. In this case it can be determined if any important alternatives were omitted.

All national guidelines are more specific and indicate that the competing alternative to which an often experimental technology is compared should be an alternative strategy that is used to treat the same condition. The Canada and Ontario guidelines state that the most commonly used and least expensive health care alternative should act as the comparator. The Australian guidelines indicate that the comparator that is most likely to be replaced has to be used in the analysis. The UK guidelines state only to justify the choice of the comparator. The Dutch guidelines indicate to use the standard option that is defined as the option of first choice in day to day practice of which the effectiveness has been proven.

First, we studied the choice of the competing alternative regarding diagnostic technologies. A literature review regarding the economic evaluation of diagnostic technologies showed that 90% of the studies compared test vs. another test and test vs. no-test (Severens & van der Wilt, 1999). However, specifically for the evaluation of a new diagnostic facility the question rises whether the technology aims to replace a diagnostic alternative or whether the technology is an add-on for existing diagnostic technologies. In the first situation, a straightforward test vs. test comparison is relevant. The latter situation is more complicated: specification of the appropriate diagnostic pathways and thus, the aspect of test sequence is relevant and should be studied (Wagner, 1983). Although in practise it is impossible to

empirically compare all possible test sequences separately, as was shown modelling can be helpful to investigate all theoretically possible test combinations and sequences.

Second, we studied the choice of the competing alternative regarding a therapeutic alternative of which no trial results were available. In this case, as was shown, modelling allows comparison against the alternative desired. The main difference between empirical studies and modelling studies is that empirical studies gather information, whereas the modelling studies synthesise information (often empirical data) without the aim of gathering new empirical data (Brennan & Akehurst, 1999).

All national guidelines concentrate on (pharmaco-) treatment and are sufficiently detailed regarding the choice of the competing alternative. However, no guideline makes specific remarks about the economic evaluation of diagnostic technologies. This sequence-aspect is distinctive for diagnostic technologies. Thus, regarding the evaluation of diagnostic technologies separate guidelines or perhaps addenda to existing guidelines should be considered. Such guidelines should concentrate on the fact that evaluation of a diagnostic technology should be designed in such a way that it reflects the future role of the technology: should it replace an existing diagnostic technology or be an add on to the current diagnostic practise. And if the latter is to be the fact, which sequence of the different diagnostic technologies is to be considered?

The choice of the comparator is highly relevant for the findings of an economic evaluation. A strict definition of a rule for defining the comparator, both in evaluating diagnostic and therapeutic technologies, does not seem to be useful, because the relevance is determined by the interest of the initiator of the study. Gold *et al.* (1996) state that ideally, all possible alternatives have to be identified (including a do-nothing alternative) and all of them should be studied. Normally this is practically impossible; thus, the technology under consideration has to be compared to existing practice. However, a practical problem may arise concerning how to define existing practice in the situation of a mixture of different approaches. As a possible solution they suggest to incorporate the main alternatives as separate options in the study or use the mixture as such as the competing alternative. In the situation that an evaluation is to be performed early in the life cycle of a medical technology, it is difficult to define the most relevant competing alternative. In this case both researcher and decision-maker should decide in consultation about the comparison made to target the study to the relevant policy question (Elsinga & Rutten, 1997). In our opinion, justification of the choice of the competing alternative should be made explicitly in the sense that one should determine if the criterion of choice is related to comparison with the most effective, cheapest, most frequently used, or any other possible alternative criterion.

The relevant costs and consequences: the time horizon and the perspective of a study

A full identification of the important and relevant costs and consequences of the alternatives should be provided. This will make a judgement possible of what specific costs and

consequences are appropriate to include in the analysis (Drummond *et al*, 1997). Related to this idea one should be aware that effects of health care technologies may influence resource use in the future and therefore the appropriate time horizon should be chosen. Besides this, the users of economic evaluations should ask the question whether the analysis covered all relevant perspectives.

Summarising the national guidelines, there are several aspects that influence the relevance of costs and consequences. All guidelines that have been studied emphasise that the relevance of costs and consequences to be analysed in an economic evaluation depends on the disease and type of patients studied. Costs and consequences that are not related to the specific disease should, in general, not be studied at all. Besides this, the differential approach of economic evaluations, aiming to find differences between the (often experimental) technology and its comparator, also highly influences the choice of the costs and consequences which are to be studied. Regarding the time horizon, Canada and Ontario recommend long range final end-points and, if necessary, utilisation of modelling techniques to estimate long-term costs and benefits. The Australian guidelines simply state that the time horizon should be appropriate to the disease and the UK guidelines indicate that the treatment path should be fully described. Thus, although formulated differently, the guidelines seem to agree about the issue of the time horizon. Regarding the study's perspective, the different guidelines have dissimilar opinions. The Canada, Ontario, UK and The Netherlands guidelines recommend a societal (or social) perspective and thus indicate to include medical, patient and caregiver, and productivity costs. The Australian guidelines expect the researcher to only include direct costs (medical and patient and caregiver costs), thus indicating the health care perspective. Only in the situation that productivity costs materially affect the results should these costs be included.

To study the impact of choosing the time horizon of an economic evaluation we studied different diagnostic strategies for invasive aspergillosis. Obviously, several time horizons can be used to determine the short and longer-term impact of the diagnostic technologies. The main objective of a diagnostic strategy is to indicate a patient for treatment; thus the end of the diagnostic phase can be the time frame for evaluating diagnostic alternatives. However, it was shown that extending the time horizon and thus including the therapeutic phase could have a significant impact on the study conclusions.

Studying the impact of a study's perspective on the relevant costs and consequences, we showed the difficulty in comparing results of economic evaluations in case different perspectives were used. And, as described before, not only the types of costs but also the way costs are calculated depend upon the perspective chosen.

Regarding the time horizon, it seems rather impossible to formulate the guidelines explicitly and to define the time horizon that has to be used in the economic evaluation of health technologies. However, in our opinion, the guidelines should be more directive: the researcher should be able to justify the choice of the time horizon. In principle a long-term time horizon should be chosen. If not, a short time horizon should be well motivated by indicating the robustness of the study results in case a shorter or longer time horizon was the

basis for the analyses. This can be done in accordance with the principles of sensitivity analyses. In addition, it seems useful to formulate specific diagnostic technology guidelines. In case a diagnostic technology is subject to an economic evaluation, limiting the time horizon to the diagnostic phase instead of including the therapeutic phase should be justified in the sense that a specific diagnostic outcome automatically leads to a defined therapeutic outcome. Thus, in both situations of evaluating diagnostic and therapeutic technologies, the impact of the time horizon chosen should be explored by estimating longer-term costs and consequences on the basis of the intermediate endpoints used. Again, modelling can be helpful to combine short-term costs and consequences (e.g. intermediate outcome as measured in a prospective study) and longer term costs and consequences (final outcome).

Given the difficulty in comparing results from cost analyses that are performed using different perspectives, the question remains which perspective is appropriate. Economic evaluations are often conducted with a specific policymaking motive. The institution or individual that is responsible for the eventual policy choice will determine the perspective of primary interest (Gold *et al.*, 1996). Therefore, using this decision-maker approach, perspectives like the third party payer perspective, health care perspective, provider perspective, and patients' perspective seem to be logical. However, it is more and more recognised that if the main objective of performing economic evaluations is to identify use of social resources for health care, than the most relevant perspective is the societal perspective. The analysis needs to consider both those who gain and those who pay the health care technology (Johannesson & Meltzer, 1998). The limited health care perspective, which is often incorrectly referred to as a societal perspective, is especially subject to criticism. As Johannesson (1995b) states, using the health care perspective is inconsistent with a societal approach because all costs and benefits irrespective of to whom they accrue should be determined and no theoretical foundation can be given for ignoring costs outside the health care sector.

Related to reimbursement decisions, in our opinion the use of the societal perspective is preferred. Using a different perspective can be motivated in the situations where it can be argued that costs outside the health care sector are relatively small or are expected to strengthen the result related to a limited perspective (Johannesson & Meltzer, 1998). In addition, a limited perspective can also be sufficient when costs outside the health care sector are expected not to be different between the health care technologies compared. Besides being directive in the choice of the perspective, the guidelines should be expanded, indicating to report the different types of costs, which are related to the distinguished perspectives, separately. This will enable the decision-maker to judge about the relevance of the different cost types and outcome components separately.

The accurate measurement of cost and consequences: productivity costs

Drummond *et al.* (1997) state that once the important and relevant costs and consequences have been identified, they must be measured in appropriate units. In general, randomised trials are considered to be the best method for gathering information about costs and consequences and trials are therefore used frequently as a vehicle for economic evaluation (Drummond & Davies, 1991).

The guidelines have different opinions about accurate measurement of cost and consequences. The Ontario guidelines prefer using data based on meta-analyses of randomised trials. Information from randomised trials are recommended by the Australian guidelines supplemented with additional information if desired. Considering this aspect, the guidelines of Canada, The UK, and The Netherlands are not prescriptive.

When performing prospective trials, accurate measurement of quantities for cost calculations seems to be possible. Accurate retrospective measurement can still be relevant in prospective trials because questionnaires that measure quantities retrospectively are often used in such trials. Regarding the issue of accurate measurement of costs and consequences we studied the possibility of measuring absence from work retrospectively as a basis for calculating productivity costs. The measurement of productivity costs was highlighted because there is considerable debate going on about this topic. As was shown, the recall period used when measuring absence from work should be appropriately short.

The existing guidelines give no guidance on the matter of accurate, retrospective measurement of costs and consequences because they implicitly assume that all empirical data in economic evaluations are measured prospectively. This is often not the case. For instance, in one of our trials we used a diary to measure medical consumption prospectively (Laheij *et al.*, 1998). After the follow-up of 12 months, we asked the respondent if they filled in the diary on a daily, weekly or monthly basis. It turned out that nearly all respondents registered their medical consumption retrospectively, either on a weekly or monthly basis. Adjustment of most guidelines on this point should be considered: in case quantities as a basis for cost analyses are measured retrospectively a maximum recall period should be advised. For measuring absence from work retrospectively, a recall period of two months seems to be advisable. General recall recommendations can hardly be made regarding all types of relevant quantities. For instance, compared to frequent recurring visits to the general physician by a dyspeptic complaint patient, a longer recall period can probably be used whenever measuring respondents' number of open-heart surgeries. Thus, for other retrospectively measured quantities such as medical consumption, further research is necessary.

The credible valuing of costs and consequences: productivity costs and willingness to pay

A clear description of the sources and methods of valuation of costs and consequences is obliged. The objective of valuing costs is to obtain an estimate of the worth of resources. Regarding the credible valuing of consequences, one should ask whether the appropriate type of economic evaluation was chosen (Drummond *et al.*, 1997)

The guidelines make the issue of credible valuing of costs and consequences operational in a different manner. The Australian guidelines require the use of a standard resource and cost list. Because of the limited perspective that is indicated, productivity costs and the way to handle compensating mechanisms is not described. The Netherlands guidelines indicate that a standard resource and cost list will be published mid 1999 and that these standards have to be used. The friction cost method will be recommended to calculate productivity costs. This method uses a 0.8 index to indicate the relation between absence from work and the productivity costs. The UK guidelines indicate that full costing (thus costs including overheads and depreciation of equipment and facilities) should be performed with no remarks on compensating mechanisms. The remaining guidelines are not directive in the way costs should be measured.

Related to the issue of credible valuing of costs and consequences we studied two aspects. First, we studied the credible valuing of costs, using the aspect of productivity costs, because considerable debate is going on about this topic. It was shown that the method that is used to calculate productivity costs - taking self-reported compensating mechanisms into account - influences the estimates of this type of costs. Second, regarding the credible valuing of consequences, we studied the appropriateness of the willingness-to-pay method for valuing the consequences of non-decisional diagnostic information. The validity of using this method could not be proven using Dutch respondents.

Although the studies perspectives might have been the same, Drummond *et al.* (1992) found that, comparing cross-national differences in cost-effectiveness of health care technologies, prices for resources were the aspects of the analysis that differed most between countries. Aspects that may influence the results and possible comparison of cost analyses are related to the time, place and patient population of analysis. Four types of biases can be distinguished in this respect; scale bias, methods bias, case mix bias, and site selection bias (Jacobs & Baladi, 1996). Scale bias is related to the idea that average cost of a health care program are independent of the magnitude of the program, thus assuming that the marginal costs are equal to the average costs. Methods bias can occur when allocation of fixed costs is not performed at all or not performed properly, thus underestimating the real costs involved. In case a volume parameter can be related to different intensities of resource use, and no correction has been made for case mix (for instance case mix weighted days in hospital) case mix bias might be the fact. Site selection bias might occur when costs are measured at a single site. The question is whether this site is the most or least efficient, or somewhere in between.

Adjustment for cost biases is possible; however, the most efficient way to investigate the influence of possible biases in cost estimates is to perform extensive sensitivity analyses (Jacobs & Baladi, 1996).

For calculating productivity costs, two competing methods can be found in the literature: the human capital approach and the friction cost approach. The first method denies any relationship between absence from work and compensating mechanisms. The friction cost method incorporates an index of 0.8 (chosen from a range between 0.6 to 0.9) for adjusting absence figures to actual productivity costs (Koopmanschap *et al.*, 1995). Our study shows that this index might not be sufficient. In fact, these differences in the calculations indicate methods bias. However, the way the latter should be prevented is not clear because the methods to be used to calculate productivity costs are still developing. This can be illustrated by the results from additional analyses regarding this topic. Using data from respondents working at a company, we found, as expected, that the reported compensating mechanisms are clearly related to the duration of absence.

Even in the situation where the study's perspective has clearly been defined and bias in the cost analysis has been prevented as much as possible, there will always be differences in the cost of diagnostic or therapeutic interventions between countries and even hospitals. Methods used to calculate costs should be clearly reported in a disaggregated, standardised way to make clear interpretation of the results possible (Drummond *et al.*, 1992) and in this manner guidelines could be more directive. This will make it easier to compare the results of cost analyses and provide more information on which principles and information cost estimates are based.

Uncertainty: statistical analysis of ratios

Two aspects of uncertainty can be distinguished in economic evaluations. First, the uncertainty can be related to deterministic variables, which can be explored by sensitivity analyses. Second, statistical uncertainty is related to stochastic variables that can include both cost and consequences. Besides this, the incremental cost-effectiveness ratio is subject to some degree of uncertainty. As Drummond *et al.* (1997) state: All economic evaluations do have some degree of uncertainty, imprecision, or methodological controversy. In principle, sensitivity analysis, statistical inference, or a combination of the two methods should allow for uncertainty in the costs and consequences.

All guidelines recommend the use of sensitivity analysis to investigate the robustness of the study findings regarding the variables. Different methods of sensitivity analyses are described in the literature, such as one- or more way sensitivity analysis and threshold analysis (Briggs *et al.*, 1994; Briggs & Sculpher, 1995). None of the national guidelines prescribe how to deal with stochastic uncertainty.

The methods to determine uncertainty related to the incremental cost-effectiveness ratios are still being developed (Briggs, 1999). Therefore, we studied the statistical analysis of this ratio and recommended the Fieller method.

It seems irrelevant to adjust the guidelines in a way that they prescribe the methods to be used to investigate stochastic uncertainty of the incremental ratio for the time being. However, after comparing several methods to explore uncertainty in the incremental costs-effectiveness ratio, we found that at least the difference in effectiveness should be established before any method to explore uncertainty of the incremental ratio is performed. Thus, the guidelines should be adjusted to emphasise that first of all, the difference in effectiveness has to be explored before any uncertainty analyses on the incremental ratio are performed. In case investigators did use some method to explore uncertainty related to the incremental cost-effectiveness ratios, the methods used and the related results should be described clearly.

CONCLUSION

The guidelines for (pharmaco-) economic evaluation that exist in different countries are quite generally phrased. Two years' experience with the Canada guidelines led to the conclusion that except for the perspective of the analysis, guidelines were, in many respects, adhered to and did not restrict investigators to specific methodologies or specific techniques (Baladi *et al.*, 1998). The question arises whether such guidelines should be strictly directive on all aspects. It seems hardly possible to gain consensus on all elements for the economic evaluation, because the methodological choices that have to be made are dependent on the specific purpose of the study. Besides this, as is shown in this thesis, the measurements of many elements that are the basis for the cost-effectiveness outcome are still in the developmental stage. However, as was described above, several recommendations for adjustments of some of the guidelines have been made. It seems logical that if the results of economic evaluations have to be comparable, more restrictive guidelines are necessary. It can be concluded that, for the time being, the results of different economic evaluations of health care technologies are not validly comparable, unless, after detailed exploration, the methods used are considered to be similar (Mason *et al.*, 1993). To facilitate this, it seems to be more logical to ensure that researchers report their methods in a transparent and standardised way, although more discussion and debate about the appropriate standards for reporting results of economic evaluations is still necessary (Mason & Drummond, 1995). However, in case these standards are defined, it will be possible for the users, i.e. the decision-makers to understand the principles of the underlying methods used and to validly make judgements about the costs and consequences as reported in the economic evaluation of health care technologies.

REFERENCES

Abrams P (1995). Objective evaluation of bladder outlet obstruction. *British Journal of Urology* 76 (suppl 1): 11-15.

Access and Financing Division (1998). *Guidelines for the pharmaceutical industry on preparation of submissions to the Pharmaceutical Benefits Advisory Committee: including major submissions involving economic analyses.* Commonwealth of Australia, Department of Health and Aged Care. Canberra: Australian Government Printing Office.

Agius RM, Lloyd MH, Campbell S, Hutchison P, Seaton A & Soutar CA (1994). Questionnaire for the identification of back pain for epidemiological purposes. *Occupational and Environmental Medicine* 51: 756-760.

Aisner J, Schimpff SC & Wiernik PH (1977). Treatment of invasive aspergillosis: relation of early diagnosis and treatment to response. *Annals of Internal Medicine* 86: 539-543.

Anaissie EJ, Darouiche RO, Abi-Said D, Uzun O, Mera J, Gentry LO, Williams T, Kontoyiannis DP, Karl CL & Bodey GP (1996). Management of invasive candidal infections: Results of a prospective, randomized, multicenter study of fluconazole versus amphotericin B and review of the literature. *Clinical Infectious Diseases* 23: 964-972.

Appel LJ, Steinberg EP, Powe NR, Anderson GF, Dwyer SA & Faden RR (1990). Risk reduction from low osmality contrast media. What do patients think it is worth? *Medical Care* 28: 324-334.

Asch DA, Patton JP & Hershey JC (1990). Knowing for the sake of knowing: the value of prognostic information. *Medical Decision Making* 10: 47-54.

Baladi J, Menon D & Otten N (1998). Use of economic evaluation guidelines: 2 years' experience in Canada. *Health Economics* 7: 221-227.

Balkany T (1993). A brief perspective on cochlear implants. *New England Journal of Medicine* 328: 281-282.

Barthel JS & Dale Everett E (1990). Diagnosis of Campylobacter pylori infections: The "Gold Standard" and the alternatives. *Review of infectious diseases* 12: s107-s114.

Berger M (1995). Design of prospective cost-effectiveness clinical trials: the critical role of the comparator group. *Drug Information Journal* 29: 1415-1420.

Bertera RL (1991). The effects of behavioral risks on absenteeism and health-care costs in the workplace. *Journal of Occupational Medicine* 33: 1119-1124.

Berwick DM & Weinstein MC (1985). What do patients value? Willingness to pay for ultrasound in normal pregnancy. *Medical Care* 23: 881-893.

Blade J, Lopez-Guillermo A, Rozman C, Granena A, Bruguera M, Bordas J, Cervantes F, Carreras E, Sierra J & Montserrat E (1992). Chronic systemic candidiasis in acute leukemia. *Annals of Hematology* 64: 240-244.

Boer WA de (1997). Diagnosis of Helicobacter pylori infection. Review of diagnostic techniques and recommendations for their use in different clinical settings. *Scandinavian Journal of Gastroenterology* 32: 35-42.

Boyko EJ (1994). Ruling out or ruling in disease with the most sensitive or specific test: short cut or wrong turn? *Medical Decision Making* 14: 175-179.

Breiman L, Friedman L, Olshen R & Stone CJ (1984). *Classification and regression trees.* Wadworth: Belmont CA.

Brennan A & Akehurst R (1999). Modelling in economic evaluation: What is it place?; What is its value? *Pharmacoeconomics* [in press].

Briggs AH, Sculpher M & Buxton M (1994). Uncertainty in the economic evaluation of health care technologies: the role of sensitivity analysis. *Health Economics* 3: 95-104.

Briggs AH & Sculpher M (1995). Sensitivity analysis in economic evaluation: a review of published studies. *Health Economics* 4: 355-371.

Briggs AH & Fenn P (1997a). Trying to do better than average: a commentary on 'statistical inference for cost-effectiveness ratios'. *Health Economics* 6: 491-495.

Briggs AH, Wonderling DE & Mooney CZ (1997b). Pulling cost-effectiveness analysis up by its bootstraps: a non-parametric approach to confidence interval estimation. *Health Economics* 6: 327-340.

Briggs AH (1999). Handling uncertainty in cost-effectiveness models. *Pharmacoeconomics* [in press].

Brouwer WBF, Koopmanschap MA & Rutten FFH (1997a). Productivity costs measurement through quality of life? A response to the recommendation of the Washington panel. *Health Economics* 6: 253-259.

Brouwer WBF, Koopmanschap MA & Rutten FFH (1997b). Productivity costs in cost-effectiveness analysis: numerator or denominator: a further discussion. *Health Economics* 6: 511-514.

Burdorf A, Post W & Bruggeling T (1996). Reliability of a questionnaire on sickness absence with specific attention to absence due to back pain and respiratory complaints. *Occupational and Environmental Medicine* 53: 58-62.

Bytzer P, Moller Hansen J & Schaffalitzky de Muckadell OB (1994). Empirical H2-blocker therapy or prompt endoscopy in management of dyspepsia. *The Lancet* 343: 811-816.

Caillot D, Casanovas O & Bernard A (1997). Improved management of invasive pulmonary aspergillosis in neutropenic patients using early thoracic computed tomographic scan and surgery. *Journal of Clinical Oncology* 15: 139-147.

Canadian Coordinating Office for Health Technology Assessment (CCOHTA) (1996). *A guidance document for the costing process.* Ottawa: CCOHTA.

Canadian Coordinating Office for Health Technology Assessment (CCOHTA) (1997). *Guidelines for economic evaluation of pharmaceuticals: Canada.* Ottawa: CCOHTA.

Chaudhary MA & Stearns SC (1996). Estimating confidence intervals for cost-effectiveness ratios: an example from a randomized trial. *Statistics in Medicine* 15: 1447-1458.

Chestnut LG, Keller LR, Lambert WE & Rowe RD (1996). Measuring heart patients' willingness to pay for changes in angina symptoms. *Medical Decision Making* 16: 65-77.

Chopra R, Strang BJ, Cervi P, Patterson KG & Goldstone AH (1991). Liposomal amphotericin B (Ambisome) in the treatment of fungal infections in neutropenic patients. *Journal of Antimicrobia and Chemotherapy.* 28: 93-104.

Chopra R, Fielding A & Goldstone AH (1992). Successful treatment of fungal infections in neutropenic patients with liposomal amphotericin (Ambisome)- A report on 40 cases from a single centre. *Leukemia and Lymphoma.* 7: 73-77.

Chouaid C, Maillard D, Housset B, Febvre M, Zaoui D & Lebeau B (1993). Cost effectiveness of noninvasive oxygen saturation measurement during exercise for the diagnosis of Pneumocystis carinii pneumonia. *American Review of Respiratory Disease* 147: 1360-1363.

Coerts JA, Baker AE, Broek P van de & Brox JPL (1996). Language development by deaf children with cochlear implants. In: Johnson CE & Gilbert JHV (eds.) *Children's language.* Hillsdale: Lawrence Erlbaum Associates Inc.

Crean GP, Holden RJ, Knill-Jones RP, Beattie AD, James WB, Majoribanks FM & Spiegelhalter DJ (1994). A database on dyspepsia. *Gut* 35: 191-202.

Crevel R van, Ven AJAM van de, Meis JFGM & Kullberg BJ (1997). Acute pulmonale histoplasmose als importziekte. *Nederlands Tijdschrift voor Geneeskunde* 141: 1242-1244 [in Dutch].

Cutler AF, Havstad S, Ma CK, Blaser MJ, Perez-Perez GJ & Schubert TT (1995). Accuracy of invasive and noninvasive tests to diagnose Helicobacter pylori infection. *Gastroenterology* 109: 136-141.

Davis A, Fortnum H & Donoghue G (1995). Children who could benefit from a cochlear implant: a European estimate of projected numbers, cost and relevant characteristics. *International Journal of Pediatric Otorhinolaryngology* 31: 221-233.

Denning DW, Evans EG, Kibbler CC, Richardson MD, Roberts MM, Rogers MM, Warnock DW & Warren RE (1997). Guidelines for the investigation of invasive fungal infections in haematological malignancy and solid organ transplantation. *European Journal of Clinical Microbiology & Infectious Diseases* 16: 424-436.

Denning DW (1994). Treatment of invasive aspergillosis. *Journal of Infectious Diseases* 28 (suppl 1): 25-33.

Diamond GA (1992). Clinical epistemology of sensitivity and specificity. *Journal of Clinical Epidemiology* 45: 9-13.

Diener A, O'Brien BJ & Gafni A (1998). Health care contingent valuation studies: a review and classification of the literature. *Health Economics* 7: 307-312.

Donaldson C, Shackley P, Abdalla M & Miedzybrodzka Z (1995). Willingness to pay for antenatal carrier screening for cystic fibrosis. *Health Economics* 4: 439-452.

Doubilet PM & Cain KC (1985). The superiority of sequential over simultaneous testing. *Medical Decision Making* 5: 447-451.

Drummond MF, Stoddart GL & Torrance GW (1987). *Methods for the economic evaluation of health care programmes.* Oxford: Oxford University Press.

Drummond MF & Davies L (1991). Economic analysis alongside clinical trials. *International Journal of Technology Assessment in Health Care* 7: 561-573.

Drummond MF, Bloom BS, Carrin G, Hillman AL, Hutchings HC, Knill-Jones RP, Pouvourville Gd & Torfs K (1992). Issues in the cross-national assessment of health technology. *International Journal of Technology Assessment in Health Care* 8: 671-682.

Drummond MF, O'Brien BJ, Stoddart GL & Torrance GW (1997). *Methods for the economic evaluation of health care programmes.* 2nd edition. Oxford: Oxford Medical Publications.

Efron B (1987). Better Bootstrap confidence intervals. *Journal of the American Statistical Association* 82: 171-188.

Efron B & Tibshirani RJ (1993). *An introduction to the bootstrap.* New York: Chapman and Hall.

Einstein DM, Herts BR, Weaver R, Obuchowski N, Zepp R & Singer A (1995). Evaluation of renal masses detected by excretory urography: cost-effectiveness of sonography versus CT. *American Journal of Roentgenology* 164: 371-375.

Elixhauser A (1993a). Health care cost-benefit and cost effectiveness analysis (CBA/CEA). From 1979 to 1990: A bibliography. *Medical Care* 31: JS 1-JS 149.

Elixhauser A, Luce BR, Taylor WR & Reblando J (1993b). Health care CBA/CEA: An update on the growth and composition of the literature. *Medical Care* 31: JS 1-JS 11.

Ellis M, Spence D, Pauw B de, Meunier F, Marinus A, Collette L, Sylvester R, Meis J, Boogaerts M, Selleslag D, Kremery V, Sinner W & MacDonald P, Doyen C & Vandercam (1998). An EORTC international multicenter randomized trials (EORTC number 19923) comparing two dosaged of liposomal amphotericin B for treatment of aspergillosis. *Clinical Infectious Diseases* 27: 1406-1412.

Ellis ME, Halim MA, Spence D, Ernst P, Clink H, Kalin M, Baillie F & Greer W (1995). Systemic amphotericin B versus fluconazole in the management of antibiotic resistant neutropenic fever-preliminary observations from a pilot, exploratory study. *Journal of Infection* 30: 141-146.

Elsinga E & Rutten FFH (1997). Economic evaluation in support of national health policy: the case of The Netherlands. *Social Science & Medicine* 45: 605-620.

Elstein AS, Kleinmuntz B, Rabinowitz M, McAuly R, Murakami J, Heckerling PS & Dod JM (1993). Diagnostic reasoning of high- and low-domain-knowledge clinicians. *Medical Decision Making* 13: 21-29.

Erkel AR van, Rossum AB van, Bloem JL, Kievit J & Pattynama PM (1996). Spiral CT angiography for suspected pulmonary embolism: a cost-effectiveness analysis. *Radiology* 201: 29-36.

Fendrick AM, Chernew ME, Hirth RA & Bloom BS (1995). Alternative management strategies for patients with suspected peptic ulcer disease. *Annals of Internal Medicine* 123: 260-268.

Flowers CR, Garber AM, Bergen MR & Lenert LA (1997). Willingness to pay utility assessment: feasibility of use in normative patient decision support systems. *Proceedings of the American Medical Informatics Association Annual Fall Symposium* 223-227.

Fryback DG & Thornbury JR (1991). The efficacy of diagnostic imaging. *Medical Decision Making* 11: 88-94.

George MJ, Snydman DR, Werner BG, Griffith J, Falagas ME, Dougherty NN & Rubin RH (1997). The independent role of cytomegalovirus as a risk factor for invasive fungal disease in orthotopic liver transplant recipients. *American Journal of Medicin* 103: 106-113.

Glasziou PP & Hilden JH (1986). Decision tables and logic in decision analysis. *Medical Decision Making* 6: 154-160.

Glasziou PP (1994). Decision tables - an underutilized tool? *Medical Decision Making* 14: 207

Glick HG (1995). Strategies for economic assessment during the development of new drugs. *Drug Information Journal* 29: 1391-1403.

Golan EH & Shechter M (1993). Contingent valuation of supplemental health care in Israel. *Medical Decision Making* 13: 302-310.

Gold MR, Siegel JE, Russell LB & Weinstein MC (1996). *Cost-effectiveness in health and medicine.* New York: Oxford University Press.

Goldberg Kahn B, Healy JC & Bishop JW (1997). The cost of diagnosis: a comparison of four different strategies in the workup of solitary radiographic lung lesions. *Chest* 111: 870-876.

Goldman L, Gordon DJ, Rifkind BM, Hulley SB, Detsky AS, Goodman DW, Kinosian B & Weinstein MC (1992). Cost and health implications of cholesterol lowering. *Circulation* 85: 1939-1941.

Goodman JL, Winston DJ, Greenfield RA, Chandrasekar PH, Fox B, Kaizer H, Shadduck RK, Shea TC, Stiff P & Friedman DJ (1992). A controlled trial of fluconazole to prevent fungal infections in patients undergoing bone marrow transplantation. *New England Journal of Medicine* 326: 845-851.

Goossens MEJB, Rutten-van Mölken MPMH, Kole-Snijders AMJ, Vlaeyen JWS, Breukelen G van & Leidl R (1998). Health economic assessment of behavioural rehabilitation in chronic low back pain: a randomised clinical trial. *Health Economics* 7: 39-51.

Groll AH, Shah PM, Mentzel C, Schneider M, Just-Nuebling G & Huebner K (1996). Trends in the postmortem epidemiology of invasive fungal infections at a University Hospital. *Journal of Infectious Diseases* 33: 23-32.

Gurney JW & Conces DJ (1996). Pulmonary histoplasmosis. *Radiology* 199: 297-306.

Harris JP, Anderson JP & Novak P (1995). An outcomes study of cochlear implants in deaf patients. *Archives of Otolaryngology: Head & Neck Surgery* 121: 398-404.

Henschke CI, Yankelevitz DF & Sicherman N (1997). Evaluation of algorithms for the diagnosis of pulmonary embolism. *Seminars in Ultrasound, CT and MR* 18: 376-382.

Hiemenz JW, Lister J & Anaissie EJ (1995). Emergency-use amphotericin B lipid complex (ABLC) in the treatment of patients with aspergillosis: historical-control comparison with amphotericin B [abstract no 3383]. *Blood* 86: 849a

Hiemenz JW & Walsh TJ (1996). Lipid formulations of amphotericin B: recent progress and future directions. *Clinical Infectious Diseases* 22: S 133-S 144.

Hiemenz JW, Cagnoni P & Tong K (1998) A cost effectiveness study comparing ambisome (liposomal amphotericin B) versus amphotericin B deoxycholate in the empirical treatment of persistently febrile neutropenic patients. Focus on Fungal Infections, Orlando, Florida [abstract].

Houston S (1994). Histoplasmosis and pulmanory involvement in the tropics. *Thorax* 49: 598-601.

Hout BA van, Al MJ, Gordon GS & Rutten FFH (1994). Costs, effects and c/e-ratios alongside a clinical trial. *Health Economics* 3: 309-319.

Hutton J, Politi C & Seeger T (1995). Cost-effectiveness of cochlear implantation of children. *Advances in Otorhinolaryngology* 50: 201-206.

Jacobs P, Bachynsky J & Baladi J (1995). A comparative review of pharmacoeconomic guidelines. *Pharmacoeconomics* 8: 182-189.

Jacobs P & Baladi J (1996). Biases in cost masurements for economic evaluation studies in health care. *Health Economics* 5: 525-529.

Jensen DM (1988). Economic assessment of peptic ulcer disease treatments. *Scandinavian Journal of Gastroenterology* 23: 214-224.

Johannessen T, Petersen H, Kleveland PM, Dybdahl JH, Sandvik AK, Brenna E & Waldum H (1990). The predictive value of history in dyspepsia. *Scandinavian Journal of Gastroenterology* 25: 689-697.

Johannesson M (1995a). On the estimation of cost-effectiveness ratios. *Health Policy* 31: 225-229.

Johannesson M (1995b). A note on the depreciation of the societal perspective in economic evaluation of health care. *Health Policy* 33: 59-66.

Johannesson M, O'Conor RM, Kobelt-Nguyen G & Mattiasson A (1997). Willingness to pay for reduced incontinence symptoms. *British Journal of Urology* 80: 557-562.

Johannesson M & Meltzer D (1998). Some reflections on Cost-effectiveness Analysis. *Health Economics* 7: 1-7.

Johnson JE, Kabler JD, Gourley MF, Dodge RW, Golubjatnikov R, Davis JP, Wheat LJ & Janzen DH (1988). Cave-associated Histoplasmosis - Costa Rica. *Archives of Dermatology* 124: 994

Jolleys JV, Donovan JL, Nanchahal K, Peters TJ & Abrams P (1994). Urinary symptoms in the community: how bothersome are they? *British Journal of Urology* 74: 551-555.

Jones S, Casswell S & Zhang JF (1995). The economic costs of alcohol-related absenteeism and reduced productivity among the working population of New Zealand. *Addiction* 90: 1455-1461.

Jönsson B & Carlsson P (1991). The effects of cimetidine on the cost of ulcer disease in Sweden. *Social Science & Medicine* 33: 275-282.

Jönsson B & Karlsson G (1996). Economic evaluation in gastrointestinal disease. *Scandinavian Journal of Gastroenterology* 31: 44-51.

Kassirer JP (1989). Our stubborn quest for diagnostic certainty; a cause of excessive testing. *New England Journal of Medicine* 320: 1489-1491.

Kent KC, Kuntz KM, Patel MR, Kim D, Klufas RA, Whittemore AD, Polak JF, Skillman JJ & Edelman RR (1995). Perioperative imaging strategies for carotid endarterectomy. An analysis of morbidity and cost-effectiveness in symptomatic patients. *Journal of the American Medical Association* 274: 888-893.

Knottnerus JA (1992a). Application of logistic regression to the analysis of diagnostic data: exact modelling of a probability tree of multiple binary variables. *Medical Decision Making* 12: 93-108.

Knottnerus JA & Leffers P (1992b). The influence of referral patterns on the characteristics of diagnostic tests. *Journal of Clinical Epidemiology* 45: 1143-1154.

Kobelt G (1997). Economic considerations and outcome measurement in urge incontinence. *Urology* 50: 100-107.

Kolts BE, Joseph B, Achem SR, Bianchi T & Monteiro C (1993). Helicobacter pylori Detection: A Quality and Cost Analysis. *American Journal of Gastroenterology* 88: 650-655.

Koopmanschap MA & Rutten FFH (1993). Indirect costs in economic studies; confronting the confussion. *Pharmacoeconomics* 4: 446-454.

Koopmanschap MA, Rutten FFH, Ineveld van BM & Roijen L van (1995). The friction cost method for measuring indirect costs of disease. *Journal of Health Economics* 15: 171-189.

Koopmanschap MA & Rutten FFH (1996a). Indirect costs. The consequence of production loss or increased costs of production. *Medical Care* 34: DS59-DS68.

Koopmanschap MA & Rutten FFH (1996b). A practical guideline for calculating indirect costs of disease. *Pharmacoeconomics* 10: 460-466.

Krabbe PF, Essink-Bot ML & Bonsel GJ (1997). The comparability and reliability of five health-state valuation methods. *Social Science & Medicine* 45: 1641-1652.

Krüger W, Stockschläder M, Rüssmann B, Berger C, Hoffknecht M, Sobottka I, Kohlschütter B, Kroschke G, Kröger N, Horstmann M, Kabisch H & Zander AR (1995). Experience with liposomal amphotericin-B in 60 patients undergoing high-dose therapy and bone marrow or peripheral blood stem cell transplantation. *British Journal of Haematology* 91: 684-690.

Kuhlman JE, Fishman EK & Siegelman SS (1985). Invasive pulmonary aspergillosis in acute leukemia: characteristic findings on CT, the CT halo sign and the role of CT in early diagnosis. *Radiology* 157: 611-614.

Lachs MS, Nachamkin I, Edelstein PH, Goldman J, Feinstein AR & Schwartz JS (1992). Spectrum bias in the evaluation of diagnostic tests: lessons from the rapid dipstick test for urinary tract infection. *Annals of Internal Medicine* 117: 135-140.

Laheij RJF, Boer WA de, Jansen JBMJ, Lier HJJ van, Sneeberger PM & Verbeek ALM. Evaluation of diagnostic performance of biopsy-based methods for determination of helicobacter pylory infection without a reference standard. [submitted].

Laheij RJF, Severens JL, Lisdonk EH van de, Verbeek ALM & Jansen JBMJ (1998). Randomised controlled clinical trial of omeprazole or endoscopy in patients with persistent dyspepsia; a cost-effectiveness analysis. *Alimentary Pharmacology and Therapeutics* 12: 1249-1256.

Landis JR & Koch GG (1977). The measurement of observer agreement for categorical data. *Biometrics* 33: 159-174.

Langley PC (1996). The November 1995 revised Australian guidelines for the economic evaluation of pharmaceuticals. *Pharmacoeconomics* 9: 341-352.

Laska EM, Meisner M & Siegel C (1997a). Statistical inference for cost-effectiveness ratios. *Health Economics* 6: 229-242.

Laska EM, Meisner M & Siegel C (1997b). The usefulness of average cost-effectiveness ratios. *Health Economics* 6: 497-504.

Lea AR (1991). *Cochlear Implants.* Australian Institute of Health, Health Care Technology Series No. 6. Canberra: Australian Government Publishing Service.

Lea AR & Hailey DM (1995). The cochlear implant; a technology for the profound deaf. *Medical Progress through Technology* 21: 47-52.

Lindholm LA, Rosen ME & Stenbeck ME (1997). Determinants of willingness to pay taxes for a community-based prevention programme. *Scandinavian Journal of Social Medicine* 25: 126-135.

Luce BR & Brown RE (1995). The use of technology assessment by hospitals, health maintenance organizations, and third party payers in the united states. *International Journal of Technology Assessment in Health Care* 11: 79-92.

Luce BR, Manning WG, Siegel JE & Lipscomb J (1996). Estimating costs in cost-effectiveness analysis. In: Gold MR, Siegel JE, Russell LB & Weinstein MC (eds.) *Cost-effectiveness in health and medicine.* New York: Oxford University Press.

Manfredi R, Mazzoni A, Nanetti A & Chiodo F (1994). Histoplasmosis capsulati and duboisii in Europe: The impact of the HIV pandemic, travel and immigration. *European Journal of Epidemiology* 10: 675-681.

Mason J, Drummond MF & Torrance G (1993). Some guidelines on the use of cost effectiveness league tables. *British Medical Journal* 306: 570-572.

Mason J & Drummond MF (1995). Reporting guidelines for economic studies. *Health Economics* 4: 85-94.

McConnel JD, Barry MJ, Bruskewitz RC, Bueschen AJ, Denton SE, Holtgrewe HL, Lange JL, McClennan BL, Mebust WK, Reilly NJ, Roberts RG, Sacks SA & Wasson JH (1994). *Benign prostatic hyperplasia: diagnosis and treatment.* Clinical practice guideline 8, No. 94-0582. Rockville: Agency for Health Care Policy and Research, Public Health Service, US Department of Health and Human Services.

McDaniels TL, Kamlet MS & Fischer GW (1992). Risk perception and the value of safety. *Risk Analysis* 12: 495-503.

McNeil BJ & Pauker SG (1984). Decision analysis for public health: principles and illustrations. *Annual Review of Public Health* 5: 135-161.

Michel BC, Seerden RJ, Rutten FFH, Beek EJR & Büller HR (1996). The cost-effectiveness of diagnostic strategies in patients with suspected pulmonary embolism. *Health Economics* 5: 307-318.

Moog JS & Geers AE (1995). Impact of the cochlear implant on the educational setting. *Advances in Otorhinolaryngology* 50: 174-176.

Mushlin AI, Mooney C, Grow V & Phelps CE (1994). The value of diagnostic information to patients with suspected Multiple Sclerosis. *Archives of Neurology* 51: 67-72.

National Agency for Medical Development and Evaluation (ANDEM) (1994). *Cochlear implants in prelingually deaf children* La Loupe. SAGER. [in French with English summary].

National Health Insurance Board (Ziekenfondsraad) (1999) *Guidelines for pharmaco-economic research.* Amstelveen: Ziekenfondsraad [in Dutch].

Nemunaitis J, Shannon-Dorcy K, Appelbaum FR, Meyers J, Owens AX, Day R, Ando D, O'Neill C, Buckner D & Singer J (1993). Long-term follow-up of patients with invasive fungal disease who received adjunctive therapy with recombinant human macrophage colony-stimulating factor. *Blood* 82: 1422-1427.

Ng TTC & Denning DW (1995). Liposomal amphotericin B (AmBisome) therapy in invasive fungal infections: evaluation of United Kingdom compassionate use data. *Archives of Internal Medicine* 155: 1093-1098.

Noel M, Levenes H, Duval P, Barbe C, Ramognino P & Verhaeghen F (1995). Epidemic of pulmonary histoplasmosis after visiting a cave in New Caladonia. *Sante* 5: 219-225 [in French with English summary].

Norman RW, Nickel JC, Fish D & Pickett SN (1994). 'Prostate-related symptoms' in Canadian men 50 years of age or older: prevalence and relationships among symptoms. *British Journal of Urology* 74: 542-550.

Nyrén O, Adami H, Gustavsson S, Lööf L & Nyberg A (1985). Social and economic effects of non-ulcer dyspepsia. *Scandinavian Journal of Gastroenterology* 20: 41-45.

O'Brien BJ, Drummond MF, Labelle RJ & Willan A (1994). In search of power and significance: issues in the design and analysis of stochastic cost-effectiveness studies in health care. *Medical Care* 32: 150-163.

O'Brien BJ & Gafni A (1996). When do the "Dollars" make sense? Toward a Conceptual framework for contingent Valuation Studies in health care. *Medical Decision Making* 16: 288-299.

O'Brien BJ, Goeree R, Torrance GW, Pauly MV, Erder H, Rusthoven J, Weeks J, Cahill M & LaMont B (1998). Assessing the value of a new pharmaceutical. A feasibility study of contingent valuation in managed care. *Medical Care* 36: 370-384.

Oravcová E, Mistrik M, Sakalová A, Drgona L, Kollár T, Helpianska L, Ilavská I, Sorkovská D, Spánik S, Kukucková E & Krcméry V (1995). Amphotericin B lipid complex to treat invasive fungal infections in cancer patients: Report of efficacy and safety in 20 patients. *Chemotherapy* 41: 473-476.

Osterhaus J, Gutterman L & Plachetka JR (1992). Healthcare resource and lost labour costs of migraine headache in the US. *Pharmacoeconomics* 2: 67-76.

Oyen WJG, Claessens RAMJ, Raemaekers JMM, Pauw B de, Meer JWM van der & Corstens FHM (1992). Diagnosing infection in febrile granulocytopenic patients with indium-111-labeled human immunoglobulin G. *Journal of Clinical Oncology* 10: 61-68.

Pascual B, Ayestaran A, Montoro JB, Oliveras J, Estibalez A, Julia A & Lopez A (1995). Administration of lipid-emulsion versus conventional amphotericin in patients with neutropenia. *Annals of Pharmacotherapy* 29: 1197-1201.

Patel R, Portela D, Badley AD, Harmsen WS, Larson-Keller JJ, Ilstrup DM, Keating MR, Wiesner RH, Krom RA & Paya CV (1996). Risk factors of invasive candida and non-candida fungal infections after liver transplantation. *Transplantation* 62: 926-934.

Pauker SG & Kassirer JP (1980). The threshold approach to clinical decision making. *New England Journal of Medicine* 302: 1109-1117.

Phelps CE (1997). Good technologies gone bad: how and why the cost-effectiveness of a medical intervention changes for different populations. *Medical Decision Making* 17: 107-117.

Poirot JL, Isnard F, Lesage S & Tabouret M (1996). *Detection of Aspergillus galactomannan in sera from hematological patients by ELISA.* 3rd Meeting of the European Confederation of Medical Mycology. Lissabon [abstract].

Polsky D, Glick HA, Willke R & Schulman K (1997). Confidence intervals for cost-effectiveness ratios: a comparison of four methods. *Health Economics* 6: 243-252.

Prentice HG, Hann IM, Herbrecht R, Aoun M, Kvaloy S, Catovsky D, Pinkerton CR, Schey SA, Jacobs F, Oakhill A, Stevens RF, Darbyshire PJ & Gibson BES (1997). A randomized comparison of liposomal versus conventional amphotericin B for the treatment of pyrexia of unknown origin in neutropenic patients. *British Journal of Haematology* 98: 711-718.

Putterman C & Ben-Chetrit E (1995). Testing, testing, testing ... *New England Journal of Medicine* 333: 1208-1211.

Pym B, Sandstad J, Byth K, Middleton WRJ & Piper DW (1990). Cost-effectiveness of cimetidine maintenance therapy in chronic gastric and duodenal ulcer. *Gastroenterology* 99: 27-35.

Raab SS & Hornberger J (1997). The effect of a patient's risk-taking attitude on the cost effectiveness of testing strategies in the evaluation of pulmonary lesions. *Chest* 111: 1583-1590.

Rabeneck L & Graham DY (1997). Helicobacter pylori: When to test, when to treat. *Annals of Internal Medicine* 126: 315-316.

Raphael K (1987). Recall bias: a proposal for assessment and control. *International Journal of Epidemiology* 16: 167-170.

Richard C, Romón I, Baro J, Insunza A, Loyola I, Zurbano F, Tapia M, Iriondo A, Conde E & Zubizarreta A (1993). Invasive pulmonary aspergillosis prior to BMT in acute leukemia patients does not predict a poor outcome. *Bone Marrow Transplantation.* 12: 237-241.

Ried W (1994). Willingness to pay for diagnostic services. A new approach to modelling patient benefits in health care. *Health Economics* 3: 255-266.

Roberts S (1993). Cochlear implants in Europe: costs and benefits. *Advances in Otorhinolaryngology* 48: 274-276.

Rohrlich P, Sarfati J, Mariani P, Duval M, Carol A, Saint-Martin C, Bingen E, Latge JP & Vilmer E (1996). Prospective sandwich ELISA galactomannan assay: early predictive value and clinical use in invasive aspergillosis. *Pediatric Infectious Disease Journal* 15: 232-237.

Roijen L van, Essink-Bot ML, Koopmanschap MA, Bonsel GJ & Rutten FFH (1996). Labor and health status in economic evaluation of health care. The health and labor questionnaire. *International Journal of Technology Assessment in Health Care* 12: 405-415.

Roosmalen MS van, Severens JL, Meis JFGM, Lees E, Barton R & Verweij PE (1998). Prevalence of antibodies to Histoplasma Capsulatum among Dutch speleologists. *Journal of Infection* 37: 200-201.

Rosier PF, Wildt MJ de, Wijkstra H, Debruyne FF & Rosette JJ de la (1996). Clinical diagnosis of bladder outlet obstruction in patients with benign prostatic enlargement and lower urinary tract symptoms: development and urodynamic validation of a clinical prostate score for the objective diagnosis of bladder outlet obstruction *Journal of Urology* 155: 1649-1654.

Rubin RH & Fischman AJ (1996). Radionuclide imaging of infection in the immunocompromised host. *Clinical Infectious Diseases* 22: 414-422.

Rutten FFH, Ineveld BM van, Ommen R van, Hout BA van & Huijsman R (1993). *Cost analysis in health care research; practical guidelines.* Utrecht: Uitgeverij Jan van Arkel [in Dutch].

Sacks JJ, Ajello L & Crockett LK (1986). An outbreak and review of cave-associated histoplasmosis capsulati. *Journal of Medical and Veterinary Mycology* 24: 313-327.

Schouw YT van der, Dijk R van & Verbeek ALM (1995). Problems in selecting the adequate patient population from existing data files for assessment studies of new diagnostic tests. *Journal of Clinical Epidemiology* 48: 417-422.

Severens JL, Donnelly JP, Meis JFGM, Vries Robbé PF de, Pauw BE de & Verweij PE (1997). Two strategies for managing invasive aspergillosis: a decision analysis. *Clinical Infectious Diseases* 25: 1148-1154.

Severens JL, Prahl C, Kuijpers-Jagtman AM & Prahl B (1998). Short term cost-effectiveness of presurgical orthopaedic treatment in children with complete unilateral cleft lip and palate. *Cleft Palate-Craniofacial Journal* 35: 222-226.

Severens JL, Vries Robbé PF de & Verbeek ALM (1999a). Optimisation of diagnostic test sequence: the probability modifying plot. *Methods of Information in Medicine* 38: 50-55.

Severens JL & Wilt GJ van der (1999b). Economic evaluation of diagnostic tests: a review of published studies. *International Journal of Technology Assessment in Health Care* 15: 480-496.

Snik AFM, Vermeulen AM, Brokx JPL, Beijk C & Broek P van de (1997a). Speech perception performance of children with a cochlear implant compared to that of children with conventional hearing aids. Part 1: The "Equivalent hearing loss concept" *Acta Otolaryngology* 117: 750-754.

Snik AFM, Vermeulen AM, Geelen CP, Brokx JPL & Broek P van de (1997b). Speech perception performance of children with a cochlear implant compared to that of children with conventional hearing aids. Part 2: Results of prelingually deaf children *Acta Otolaryngology* 117: 755-759.

Sonnenberg A & Everhart JE (1997). Health impact of peptic ulcer in the United States. *American Journal of Gastroenterology* 92: 614-620.

Sox HC, Blatt MA, Higgins MC & Marton KI (1988). Selecting and interpretation of diagnostic tests. In: Sox HC, Blatt MA, Higgings MC & Marton KI (eds.) *Medical decision making.* Boston: Butterworths.

Spilker B (1996). *Quality of life and pharmacoeconomics in clinical trials.* Philadelphia: Lippincott-Raven Publishers.

Statistics Netherlands (Centraal Bureau voor de Statistiek, CBS) (1997a). *Vadecum of health statistics of the Netherlands 1997.* 's-Gravenhage: SDU uitgeverij [in Dutch].

Statistics Netherlands (Centraal Bureau voor de Statistiek, CBS) (1997b). *Annual Statistics 1997.* Voorburg/Heerlen: CBS [in Dutch].

Statistics Netherlands (Centraal Bureau voor de Statistiek, CBS) (1999). *Annual Statistics 1999.* Voorburg/Heerlen: CBS [in Dutch].

Stynen D, Goris A, Sarfati J & Latge JP (1995). A new sensitive sandwich enzyme-linked immunosorbent assay to detect galactofuran in patients with invasive aspergillosis. *Journal of Clinical Microbiology* 33: 497-500.

Sulahian A, Tabouret M, Ribaud J, Sarfati J, Gluckman E, Latge JP & Derouin F (1996). Comparison of an enzyme immunoassay and latex agglutination test for detection of galactomannan in the diagnosis of invasive aspergillosis. *European Journal of Clinical Microbiology & Infectious Diseases* 15: 139-145.

Summerfield AQ & Marshall DH (1995a). Cost-effectiveness of cochlear implantation. In: Anonymous *Cochlear implantation for the UK 1990-1994: report by the MRC Institute of Hearing Research on the evaluation of the National Cochlear Implant Programme.* Nottingham: MRC Institute of Hearing Research.

Summerfield AQ, Marshall DH & Davis AC (1995b). Cochlear implantation: demand, costs and utility. *Annals of Otology, Rhinology and Laryngology* Suppl. 166: 245-248.

Suzaki A, Kimura M, Kimura S, Shimada K, Miyaji M & Kaufman L (1995). An outbreak of acute pulmonary histoplasmosis among travelers to a bat-inhabited cave in Brazil *Kansenshogaku-Zasshi* 69: 444-449 [in Japanese with English summary].

Swanink CMA, Meis JFGM, Rijs AJMM, Donnelly JP & Verweij PE (1997). Specificity of a sandwich enzyme-linked immunosorbent assay for detecting Aspergillus galactomannan. *Journal of Clinical Microbiology* 35: 257-260.

Tabone MD, Vu Thien H, Latge JP, Landman-Parker J & Leverger G (1996). *Galactomannan detection by sandwich enzyme-linked immunosorbent assay in the diagnosis and follow-up of invasive aspergillosis.* 3rd Meeting of the European Confederation of Medical Mycology. Lissabon [abstract].

Tambour M & Zethraeus N (1998a). Bootstrap confidence intervals for cost-effectiveness ratios: some simulation results. *Health Economics* 7: 143-147.

Tambour M, Zethraeus N & Johannesson M (1998b). A note on confidence intervals in cost-effectiveness analysis. *International Journal of Technology Assessment in Health Care* 14: 467-471.

Thompson MS (1986). Willingness to pay and accept risks to cure chronic disease. *American Journal of Public Health* 76: 392-396.

Tollemar J, Ringdén O & Tydén G (1990). Liposomal amphotericin-B (Ambisome) treatment in solid organ and bone marrow transplant recipients. Efficacy and safety evaluation. *Clinical Transplantation* 4: 167-175.

Tollemar J, Andersson S, Ringdén O & Tydén G (1992). A retrospective clinical comparison between antifungal treatment with liposomal amphotericin B (Ambisome) and conventional amphotericin B in transplant recipients. *Mycoses* 35: 215-220.

Torrance GW, Siegel JE & Luce BR (1996). Framing and designing the cost-effectiveness analysis. In: Gold MR, Siegel JE, Russell LB & Weinstein MC (eds.) *Cost-effectiveness in health and medicine.* New York: Oxford University Press.

Towse A (1997). *Guidelines for the economic evaluation of pharmaceuticals: can the UK learn form Australia and Canada?* London: The Office of Health Economics

Vermeulen AM, Beijk CM, Brokx JPL, Borne SCF van de & Broek P van de (1995). Development of speech perception abilities of profoundly deaf children: a comparison between children with cochlear implants and those with conventional hearing aids. *Annals of Otology, Rhinology and Laryngology* 104: 215-217.

Verweij PE, Stynen D, Rijs AJMM, Pauw BE de, Hoogkamp-Korstanje JAA & Meis JFGM (1995). Sandwich enzyme-linked immunosorbent assay compared with Pastorex test for diagnosing invasive aspergillosis in immunocompromised patients. *Journal of Clinical Microbiology* 33: 1912-1914.

Verweij PE, Donnelly JP, Pauw BE de & Meis JFGM (1996). Prospects for the early diagnosis of invasive aspergillosis in the immunocompromised patient. *Review of Medical Microbiology* 7: 105-113.

Wagner JL (1983). The feasability of economic evaluation of diagnostic procedures. *Social Science & Medicine* 17: 861-869.

Wakker P & Klaassen MP (1995). Confidence intervals for cost/effectiveness ratios. *Health Economics* 4: 373-381.

Walsh T, Bodensteiner D, Hiemenz J, Thaler S, Greenberg RN, Arndt C, Holcenberg JS, Schwartz C, Pappas P, Dummer S, Marsh R, Schuster M & Seibel N. (1997). *A randomized, double-blind trial of ambisome (liposomal amphotericin B) versus amphotericin B in the empirical treatment of persistently febrile neutropenic patients.* 37th Interscience Conference on Antimicrobiology Agents Chemotherapy Toronto, Canada; LM 90 [Abstract].

Walsh TJ, Finberg R, Arndt Hiemenz J, Schwartz C, Bodensteiner D, Pappas P, Seibel N, Greenberg RN, Dummer S, Schuster M & Holcenberg JS (1999). Liposomal amphotericin B for empirical treatment of patients with persistent fever and neutropenia. *New England Journal of Medicine* 340: 764-771.

Weinstein MC, Siegel JE, Garber AM, Lipscomb J, Luce BR, Manning Jr WG & Torrance GW (1997). Productivity costs, time costs and health-related quality of life: a response to the Erasmus group. *Health Economics* 6: 505-510.

Willan AR & O'Brien BJ (1996). Confidence intervals for cost-effectiveness ratio's: an apllication of Fieller's theorem. *Health Economics* 5: 297-305.

Winston DJ, Chandrasekar PH, Lazarus HM, Goodman JL, Silber JL, Horowitz H, Shadduck RK, Rosenfeld CS, Ho WG & Islam MZ (1993). Fluconazole prophylaxis of fungal infections in patients with acute leukemia: results of a randomized, placebo-controlled, double blind, multicenter trial. *Annals of Internal Medicine* 118: 495-503.

Woodward RS, Schnitzler MA & Kvols LK (1998). Reduced uncertainty as a diagnostic benefit: an initial assessment of somatostatic receptor scintigraphy's value in detecting distant metastases of carcinoid liver tumours. *Health Economics* 7: 149-160.

Wyatt JR, Niparko JK, Rothman ML & Lissovoy G de (1995). Cost effectiveness of the multichannel cochlear implant. *American Journal of Otology* 16: 52-62.

SUMMARY

This thesis concentrates on some methodological issues in the economic evaluation in health care. The objective of the economic evaluations of health care technologies in general is to provide information about efficiency of competing alternatives. The methodology of economic evaluation is still evolving. However, the methodological principles of the economic evaluations that are currently executed are of influence on the study results. The following issues are discussed: the choice of the competing alternative, the relevant costs and consequences, the accurate measurement of costs and consequences, credible valuing of costs and consequences, and the uncertainty of the results of an economic evaluation.

The issue of the choice of the competing alternative is discussed using two topics (Chapters 2.1 and 2.2). First, the competing alternative when evaluating diagnostic technologies is discussed, concentrating on the problem of deciding about the optimal sequence of diagnostic tests. Test sequences can be structured in decision trees, but unmanageable bushy decision trees result when the sequence of two or more tests is investigated. Most modelling techniques include tests on the basis of gain in certainty. The aim of studying this issue was to explore a model for optimising the sequence of diagnostic tests based on efficiency criteria. The probability modifying plot shows when, in a specific test sequence, further testing is redundant and which costs are involved. In this way different sequences can be compared. The sequence of diagnostic tests was optimised on the basis of efficiency, which was either defined as the test sequence with the least number of tests or the least total cost for testing. Further research on the model is needed to handle current limitations. The model is illustrated with two applications using data on the diagnosis of Helicobacter Pylori and Bladder Outlet Obstruction (Benign Prostatic Hyperplasia)(Chapter 2.1).

Second, the competing alternative when evaluating health care technologies is discussed, concentrating on the problem of modelling therapeutic alternatives. A decision analytic model was developed to compare the effectiveness and costs of two strategies for the empirical treatment of invasive fungal infection (IFI) in patients with haematological malignancies. Empirical treatment with amphotericin B desoxycholate (DC-Amb), followed by liposomal amphotericin B (L-Amb) in case of treatment failure or nephrotoxicity (strategy DC/L-Amb) was compared to first line treatment with L-Amb (strategy L-Amb). Estimates for the probability variables for successful treatment, treatment failure and the occurrence of side effects were derived from published reports. Effectiveness was expressed as survival to hospital discharge. Drug costs and the costs associated with the treatment of side effects were determined according to the perspective of the health-care system. The incremental cost effectiveness ratio was calculated to reflect the financial effort needed to gain more effectiveness. Extensive sensitivity analyses were performed using both probability and cost

variables to analyse the impact of the variables on the study findings. The L-Amb strategy increased the probability of survival from 77% using DC/L-Amb to 85%. The expected cost of L-Amb was USD 27,810 per patient compared to USD 12,776 per patient for the DC/L-Amb strategy. The incremental cost per life saved was nearly USD 183,000. The sensitivity analysis revealed that if the cost of L-Amb per day is less than USD 122, L-Amb is more effective expressed in lives saved at a lower cost compared to the DC/L-Amb strategy. The L-Amb strategy results in a higher survival to hospital discharge. The expected costs are higher compared to DC/L-Amb strategy. Decreasing the costs of L-Amb has a positive impact on the incremental cost-effectiveness ratio. It was shown that modelling is a method that makes comparison of relevant alternatives, which have not been described in the literature, possible (Chapter 2.2).

The issue of the relevant costs and consequences in an economic evaluation of a health care technology is discussed using two separate topics (Chapters 3.1 and 3.2). First, a decision analytic model for the diagnosis of invasive aspergillosis studied the importance of defining a time horizon of analysis. A diagnostic approach was devised, based on screening plasma for an Aspergillus antigen with use of a sandwich enzyme-linked immunosorbent assay (ELISA), thoracic computed tomographic scanning (CT), and radionuclide imaging (IgG) for managing patients at risk for invasive aspergillosis. The conventional strategy relied only on the presence of clinical symptoms: persistent fever, and chest roentgenographic findings. Use of the alternative strategy reduced the number of patients who would receive antifungal treatment empirically, but this strategy was more expensive. A 13% prevalence of infection resulted in equal costs for both strategies. As much as 43.3% of the patients treated empirically could be given liposomal amphotericin B (L-AmB) before the conventional strategy became the most expensive. The costs of the alternative strategy were less than those of the conventional strategy when >5.3% of all patients, irrespective of strategy, were treated with L-AmB. The model shows the relevance of defining the time horizon when evaluating diagnostic technologies. Limiting the time horizon to the diagnostic process itself gave different results compared to a longer time horizon, thus incorporating the treatment process (Chapter 3.1).

Second, the importance of defining the study's perspective is studied (Chapter 3.2). The definition of the perspective influences the relevancy of the different costs that have to be analysed. Besides this, the choice of the perspective determines the way cost prices should be calculated. A cost analysis was performed alongside a clinical study of cochlear implants in children in The Netherlands. Between 1993 until 1996 106 deaf children were screened as candidates for a cochlear implant. Of these, 20 children were implanted. For the selection and implantation, data from the University Hospital Nijmegen were used. Data of rehabilitation and after care were obtained from the Institute for the Deaf in Sint Michielsgestel. Real costs of medical care were calculated, using a societal perspective and a time horizon of five years. Volumes of utilization of human resources and materials were prospectively registered during the follow up of one year. For the subsequent period volumes were modeled on the basis of

planned after care activities. Basis for the calculations were 1994 prices. The economic consequences of cochlear implants on educational needs were not taken into account because of the limited period of follow up. Total medical costs of an implanted child were Dfl. 117,617. The breakdown of the costs are as follows: costs of selection phase were Dfl. 14,256, costs of implantation, including the hardware, Dfl. 56,014, costs of rehabilitation Dfl. 24,706 and the costs of after care Dfl. 22,641. The cochlear implants hardware was a large part of the total costs (Dfl. 46,397). Non medical costs were Dfl. 3,383. Sensitivity analysis of the rate of implanted children as part of the number of screened children did not show a large impact on the total costs. Compared to the results from cost analyses of other studies, the costs of the paediatric cochlear implants program in The Netherlands are relatively high. Most difference can be explained by methodological differences. For most studies the study's perspective is not defined (Chapter 3.2).

The issue of accurate measurement of costs and consequences is illustrated by the concept of productivity costs (Chapter 4). The purpose of this study was to discuss precision and accuracy of a retrospective self-administered questionnaire on sick leave. Employees of a company were asked to indicate the number of days absent from work due to illness during the past 2 weeks, 4 weeks, 2 months, 6 months, and the past 12 months. The percentage of respondents with an absolute difference of a maximum of respectively 0, 1, 2, 3, 4, 5, 6, 7, 8, and 9 or more days between reported, and company-registered absence due to illness, the so called level of precision, was determined. A systematic difference was tested with a sign rank test. Of the reported data, 95% matched the registered data perfectly when the recall period was limited to 2 and 4 weeks. This percentage decreased to 87%, 57%, and 51% for 2 months, 6 months, and 12 months, respectively. No systematic positive or negative difference was found between registered and reported sick leave. This was confirmed by the fact that no increasing proportional errors could be found. The results suggest that the recall period for retrospective measurement of sick leave is limited according to the precision level, which seems to be appropriate for the subject and the purpose of the study. We recommend using a recall period of no more than two months.

Credible valuing of costs and consequences of health care technologies is studied regarding two aspects (Chapter 5.1 and 5.2). First, using an example in the field of gastrointestinal disease, credible valuing of productivity costs is discussed. Productivity costs are normally calculated by multiplying days absent valued by gross earnings. This, however, might lead to an overestimation because compensating mechanisms are not accounted for. A conservative approach is proposed for calculating productivity costs, taking absence-compensating mechanisms into account. Patients who visited their general practitioner for the first time with dyspeptic complaints and patients who were known to have persistent dyspeptic complaints were enrolled in two studies. In total, 136 patients completed a questionnaire about their employment situation, absence from work and absence compensating mechanisms. Sixty-six of the respondents had a paid job, of which 25 (38%) reported absence from work during the previous 4 weeks (average 3.0 days, 1.9 days related to

dyspeptic complaints). More than 50% of the employed respondents answered that colleagues could compensate for absence, and only in 8% of the cases was absence compensated for by overtime. Using our conservative approach, only one-quarter of the productivity costs remained, compared to the current approach of valuing each day absent as a loss of productivity. Using both the current and the conservative approaches, analogous to the principles of sensitivity analysis, to avoid overestimation of productivity costs seems advisable (Chapter 5.1).

Second, several alternatives can be used to value the consequences of a health care technology. In this study focus is on the construct validity of the willingness to pay (WTP) method to value non-decisional diagnostic information. Six hypotheses were tested regarding the relationship between WTP and, respectively, the subjective importance of testing, the burden of testing, the perceived reliability of the test, the subjective belief of being infected, the perceived severity of the disease, and the perceived possibility of treatment. The hypotheses were tested among individuals at increased risk of histoplasmosis using a questionnaire. Univariate and multivariate relationships were examined by chi-squared test and logistic regression, respectively. Eighty-four Dutch speleologists participated in the study, 76 of whom gave WTP information. A significant relation was found between WTP and subjective importance of testing. No significant relationship was found between WTP and the other hypotheses. The multivariate analysis showed no further significant relationships. Our results indicate that WTP measurement is not a valid method to assess the value that the respondents place on non-decisional diagnostic information. This may primary be related to the fact that being Dutch citizens the respondents were not familiar with paying health care facilities out of pocket. This would mean that WTP is not a valid method of choice when assessing the monetary value that subjects place on health care interventions in countries with comprehensive coverage schemes (Chapter 5.2).

The issue of uncertainty of the results of an economic evaluation is described in the final study (Chapter 6). Several methods have been explored to estimate the confidence interval of the incremental cost-effectiveness ratio (ICER). In this study Fieller intervals and three methods for calculating bootstrap intervals are compared. Based on trial data, 10,000 trials were simulated, resulting in 10,000 ICERs and 10,000 confidence intervals for each method. The mean of the ICERs was used as estimate for the true population ICER. The adequacy of the methods to compute confidence intervals was determined by comparing the percentage of confidence intervals containing this estimate. The impact of 'no significant difference of effectiveness' was investigated by varying the difference in effectiveness. Both Fieller and bootstrap methods lead to unsatisfactory results when the difference in effectiveness is approximately zero. In the situation where this difference is significant, the four methods for calculating confidence intervals for ICER do not give very different results, but Fieller's interval performs best. Since Fieller's confidence limits are relatively easy to compute compared to bootstrap simulations, we recommend using this method.

The methodological issues in the economic evaluation in health care that were dealt with are discussed within the framework of guidelines for (pharmaco-) economic evaluation that exist in Canada, Ontario, Australia, United Kingdom, and The Netherlands (Chapter 7). It is concluded that the guidelines are quite generally phrased. It seems hardly possible to gain consensus on all methodological aspects of economic evaluations, because, as was shown in this thesis, the methods of this type of research are still being developed. However, appropriate standards for reporting study results should ensure researchers to report their methods in a transparent and standardised way. This will enable decision-makers to validly judge about costs and consequences as reported in the economic evaluation of health care technologies.

SAMENVATTING

Dit proefschrift behandelt een aantal methodologische aspecten van economische evaluaties in de gezondheidzorg. Het doel van economische evaluaties van medische technologieën (producten, apparaten of processen) die toegepast (gaan) worden in de gezondheidszorg is hoofdzakelijk het genereren van informatie over de efficiëntie van mogelijke alternatieven. De methoden die bij dergelijk onderzoek toegepast worden zijn in ontwikkeling. Echter, de methodologische principes die momenteel gehanteerd worden zijn medebepalend voor de uitkomsten van economische evaluaties. De volgende aspecten worden behandeld: de keuze van het vergelijkingsalternatief, de relevantie van kosten en effecten, accuraatheid van het meten van kosten en effecten, betrouwbaarheid van de waardering van kosten en effecten en de onzekerheid van de resultaten van een economische evaluatie.

Het aspect de keuze van het vergelijkingsalternatief wordt aan de hand van twee onderwerpen besproken (Hoofdstukken 2.1 en 2.2). Allereerst staat de keuze centraal van het vergelijkingsalternatief bij een economische evaluatie van diagnostische technologieën waarbij de nadruk ligt op de optimale sequentie van diagnostische tests. Testsequenties kunnen worden bestudeerd met behulp van besliskundige modellen, echter, weinig inzichtelijke beslisbomen ontstaan bij het onderzoeken van de sequentie van twee of meer testen. Andere modellen includeren testen in de volgorde van afnemende winst in zekerheid omtrent de ziektetoestand van patiënten. Hoofdstuk 2.1 beschrijft een evaluatiemethode om de optimale sequentie, gericht op efficiëntie, te bepalen van diagnostische tests. Deze methode laat in grafieken zien in welke situatie een volgende diagnostische test in relatie tot een behandelingsbeslissing overbodig is, en op grond waarvan de daarmee gepaarde kosten kunnen worden bepaald. Op deze manier kunnen alle mogelijke sequenties van tests worden geanalyseerd, waarna de optimale volgorde kan worden bepaald. De methode wordt geïllustreerd aan de hand van databestanden betreffende de diagnose van Helicobacter Pylori en Benigne Prostaat Hyperplasie.

Vervolgens staat de keuze van het vergelijkingsalternatief bij een economische evaluatie van therapeutische technologieën centraal, waarbij de nadruk ligt op modellering van therapeutische alternatieven. Een besliskundig model wordt beschreven waarin de effectiviteit en kosten van twee empirische strategieën worden vergeleken voor de behandeling van patiënten verdacht van invasieve fungale infectie (IFI) bij hematologische maligniteiten. Empirische behandeling met amphotericine B desoxycholaat (DC-Amb) eventueel gevolgd door behandeling met het liposomale amphotericine B (L-Amb) wordt vergeleken met directe behandeling met L-Amb. In de literatuur is geen prospectieve studie beschreven waarin de sequentiële strategie wordt bestudeerd. Rapportage betreffende de kansen op succesvolle behandeling, falen van de behandeling, en vóórkomen van bijwerkingen van louter DC-Amb en L-Amb zijn wel in de literatuur voorhanden. In het

model wordt overleving tot ontslag uit het ziekenhuis als effectmaat gehanteerd. Kosten van medicatie en ziekenhuisverblijf zijn berekend op basis van een bedrijfseconomische perspectief. Uitgebreide sensitiviteitsanalyses laten de invloed van aannames in het model zien. Directe behandeling met L-Amb verhoogt de overlevingskans van 77% naar 85%, waarbij de kosten respectievelijk US dollar (USD) 27,810 en 12,776 zijn. De incrementele kosten per gewonnen leven komen uit op USD 183,000. De kosten van liposomale amphotericine B hebben een grote invloed op de uitkomsten. Dit hoofdstuk illustreert de mogelijkheid om vergelijking van beleidsrelevante alternatieven die niet in de literatuur beschreven worden met behulp van een modelleringstudie expliciet te analyseren (Hoofdstuk 2.2).

Het aspect van de relevantie van kosten en effecten in een economische evaluatie wordt behandeld aan de hand van twee studies (Hoofdstukken 3.1 en 3.2). Een besliskundig model beschrijft het belang van het vaststellen van de tijdshorizon van analyse bij de vergelijking van twee diagnostische strategieën voor het diagnostiseren van invasieve aspergillose. Eén strategie is gebaseerd op regelmatige controle van plasma op de aanwezigheid van aspergillus antigen (ELISA test), eventueel bevestigend door CT-scan of IgG-scan. De conventionele benadering is gebaseerd op het vóórkomen van klinische verschijnselen (persisterende koorts) en een röntgenfoto van de longen, echter deze strategie leidt tot overbehandeling: 24% van de risicopatiënten wordt behandeld terwijl slechts 4% daadwerkelijk een invasieve infectie heeft. De alternatieve benadering gebaseerd op de ELISA tests reduceerde het aantal patiënten dat in aanmerking zou komen voor anti-fungus behandeling, maar deze strategie is wat betreft de kosten van diagnostiek minder gunstig. Echter, door in het model het kostbare liposomale amphotericine B op te nemen, blijkt de ELISA strategie te prefereren. Dus door de tijdshorizon van analyse te verbreden van alleen het diagnostisch traject naar het therapeutische traject en zodoende ook de kosten van dit therapeutische traject in de analyse te betrekken worden de resultaten van de vergelijking beïnvloed. Dit illustreert het belang van een juiste bepaling van de tijdshorizon van een economische evaluatie (hoofdstuk 3.1).

Ten tweede wordt het aspect van de relevantie van kosten en effecten in een economische evaluatie bestudeerd aan de hand van het perspectief van een economische evaluatie (Hoofdstuk 3.2). De keuze van het perspectief is medebepalend voor de uitkomsten van een economische evaluatie omdat hierdoor de keuze van de kostensoorten en de manier waarop kostprijzen worden berekend wordt beïnvloed. Een kostenanalyse die onderdeel was van een klinische studie betreffende cochleaire implantaten bij kinderen wordt beschreven. Tussen 1993 en 1996 doorliepen 106 dove kinderen een selectieprocedure voor mogelijke behandeling met een cochleair implantaat. Uiteindelijk werden hiervan 20 kinderen geïmplanteerd. Voor de kosten van selectie en feitelijke implantatie werden gegevens uit het AZN St. Radboud gebruikt. Gegevens over de kosten van rehabilitatie en nazorg waren beschikbaar gesteld door het Instituut voor Doven te Sint Michielsgestel. Zogenaamde werkelijke kosten gebaseerd op een tijdshorizon van 5 jaar werden berekend. Volumes van de

inzet van menskracht en middelen waren prospectief geregistreerd gedurende selectie, implantatie en rehabilitatie in het eerste jaar na implantatie en kosten van nazorg werden op basis van geplande zorg bepaald. De totale medische kosten kwamen uit op Dfl. 117.617 per geïmplanteerd kind (selectie Dfl. 14.256, implantatie 56.014, rehabilitatie Dfl. 24.706 en nazorg Dfl. 22.641). Deze resultaten werden vergeleken met de resultaten van een aantal buitenlandse studies waaruit bleek dat de Nederlandse berekeningen hoog uit kwamen. Echter, het is onmogelijk deze resultaten zonder meer te vergelijken omdat onduidelijk is vanuit welk perspectief de buitenlandse studies zijn uitgevoerd en dus welke kostensoorten en methoden van kostenberekeningen zijn gehanteerd.

Het aspect accuraatheid van het meten van kosten en effecten wordt geïllustreerd aan de hand van het concept van de productiviteitskosten (Hoofdstuk 4). Productiviteitskosten zijn kosten gerelateerd aan arbeidsverzuim als gevolg van de gezondheidstoestand van personen. Het doel van de studie is de precisie en accuraatheid van het retrospectief meten van arbeidsverzuim met behulp van een vragenlijst te onderzoeken. In de literatuur worden herinneringsperioden tot 12 maanden gevonden. In deze studie werd medewerkers van een commercieel bedrijf gevraagd aan te geven hoeveel dagen zij zich hadden ziek gemeld in verband met hun gezondheidstoestand gedurende de laatste 2 weken, 4 weken, 2 maanden, 6 maanden en 12 maanden. Deze gegevens werden vergeleken met de gegevens uit de bedrijfsregistratie die als gouden standaard werd gehanteerd. Het percentage respondenten met een absoluut verschil tussen zelfrapportage en registratie van 0, 1, 2, 3, 4, 5, 6, 7, 8 en 9 of meer dagen werd berekend. Een systematisch verschil werd onderzocht met een rangtekentoets. Bij het hanteren van een herinneringsperiode van 2 en 4 weken rapporteerden 95% van de respondenten het juiste aantal dagen. Dit percentage neemt af tot 87%, 57% en 51% voor de langere herinneringsperioden. Statistische analyses konden geen systematische afwijking aantonen. De resultaten suggereren dat de precisie van retrospectieve meting van arbeidsverzuim afneemt met het langer worden van de herinneringsperiode op basis waarvan een herinneringsperiode van maximaal 2 maanden aan te bevelen is.

Betrouwbaarheid van de waardering van kosten en effecten werd bestudeerd aan de hand van betrouwbare waardering van productiviteitskosten en de construct validiteit van de effectiviteitmaat 'willingness-to-pay' (WTP). Productiviteitskosten als gevolg van arbeidsverzuim gerelateerd aan de gezondheidstoestand van een persoon worden normaliter berekend door de afwezigheiddagen te vermenigvuldigen met de gemiddelde salarislast voor de werkgever. Echter, deze waardering van arbeidsverzuim zou kunnen leiden tot een overschatting omdat bij deze calculatie geen rekening wordt gehouden met mogelijke compensatiemechanismen om verzuim op te vangen. In deze studie wordt een conservatieve methode voorgesteld die rekening houdt met compensatiemechanismen bij kortdurend arbeidsverzuim waarbij gebruik wordt gemaakt van gegevens uit een klinische studie betreffende patiënten met maagklachten. In de eerste studie betrof het patiënten die voor het eerst bij de huisarts kwamen met maagklachten en in een tweede studie betrof het patiënten die met aanhoudende maagklachten weer hun huisarts bezochten. In totaal vulden 136

patiënten een vragenlijst in betreffende arbeidsverzuim en compensatiemechanismen bij verzuim. Van degenen met betaald werk in deze groep (66 patiënten) rapporteerden 25 patiënten te hebben verzuimd in de laatste 4 weken. Meer dan 50% van de werkenden rapporteerden dat collegae in staat waren hun afwezigheid te compenseren zonder daarvoor te hoeven overwerken. Rekening houdend met de gerapporteerde compensatiemechanismen resteerde ongeveer 25% van de productiviteitskosten in vergelijking met de conventionele methode. Aangeraden wordt om in economische evaluatie met behulp van een sensitiviteitsanalyse het verschil tussen de conventionele en de meer conservatieve benadering te analyseren (Hoofdstuk 5.1).

De construct validiteit van WTP als uitkomstmaat van een economische evaluatie werd bestudeerd in het kader van de niet-beslissing informatie van een diagnostische test (Hoofdstuk 5.2). Diagnostische tests genereren informatie en in het geval daar geen behandelingsbeslissing op gebaseerd wordt, wordt dit ook wel de niet-beslissing informatie van een test genoemd. De vraag werd onderzocht hoe dergelijke informatie betrouwbaar gewaardeerd kan worden. Zes hypothesen werden uit de literatuur afgeleid waarmee de relatie werd beschreven tussen WTP en het subjectieve belang van een test, de belasting van een test, de veronderstelde betrouwbaarheid van een test, de veronderstelde aanwezigheid van aandoening, de veronderstelde ernst van de aandoening, en de veronderstelde mogelijkheid voor behandeling. Deze hypothesen werden getoetst in een studie bij een risicopopulatie voor een histoplasmose-infectie. Een groep van 84 Nederlandse speleologen nam deel aan de studie waarvan 76 de WTP vraag beantwoorden. Er werd een significante relatie gevonden tussen de WTP antwoorden en de antwoorden betreffende het subjectieve belang van een test. Geen enkele andere relatie kon worden aangetoond tussen de WTP-antwoorden en de antwoorden betreffende de hypothesen. Deze resultaten leiden tot de veronderstelling dat de construct validiteit van WTP als uitkomstmaat voor niet-beslissing informatie van een diagnostische test onvoldoende is. Een mogelijke verklaring hiervoor is dat de meeste Nederlanders niet gewend zijn om rechtstreeks voor voorzieningen in de gezondheidszorg te betalen.

Het aspect onzekerheid van de resultaten van een economische evaluatie wordt besproken in Hoofdstuk 6. Verschillende methoden werden vergeleken waarmee een betrouwbaarheidsinterval van een incrementele kosten-effectiviteitsratio kan worden berekend. In dit hoofdstuk werden Fieller intervallen en drie methoden voor de berekening van betrouwbaarheidsintervallen op basis van bootstrap simulaties vastgesteld waarbij gegevens uit een prospectieve studie werden gebruikt. Deze studie werd nog eens 10.000 keer gesimuleerd en hierbij werden 10.000 ratio's bepaald. Het gemiddelde hiervan werd als werkelijke populatie waarde beschouwd. Voor iedere simulatie met de bijbehorende ratio werden betrouwbaarheidsintervallen vastgesteld uitgaande van de vier verschillende methoden. De nauwkeurigheid van iedere methode werd weergegeven door het percentage van de intervallen dat de 'werkelijke' populatie ratio bevatte. Zowel de Fieller methode als de drie bootstrap gebaseerde methoden leidden tot onbevredigende resultaten in het geval het effectverschil tussen alternatieven dicht bij nul ligt. Bij een significant verschil in effectiviteit

leiden zowel Fieller als de bootstrap methoden tot acceptabele resultaten, alhoewel Fieller het meest nauwkeurig is. Aangezien deze methode in vergelijking met de bootstrap methoden relatief eenvoudig is krijgt deze de voorkeur.

Het proefschrift wordt afgesloten met een beschouwing van de behandelde methodologische aspecten van economische evaluaties in het kader van richtlijnen voor het uitvoeren van deze studies (Hoofdstuk 7). De conclusie wordt getrokken dat de richtlijnen voor (pharmaco-) economische evaluaties zoals die bestaan in Canada, Ontario, Australië, Verenigd Koninkrijk en Nederland vrij algemeen verwoord zijn. Een van de redenen hiervoor is dat de methodologie van economische evaluaties zich nog ontwikkelt. In dit proefschrift worden met betrekking tot enkele aspecten van economische evaluaties concrete suggesties gedaan. Met het doel een betere vergelijking van de resultaten van dergelijk onderzoek mogelijk te maken is het zinvol de onderzoeksmethoden van economische evaluaties in de gezondheidszorg verder te ontwikkelen en op basis hiervan meer directieve richtlijnen te formuleren.

CO-AUTHORS AND AFFILIATIONS

Th.M. de Boo, MSc
Department of Medical Statistics, University of Nijmegen, The Netherlands

J.J. Bos, MSc
Department of Medical Technology Assessment, University of Nijmegen, The Netherlands

P. van den Broek, MD PhD
Department of ENT, University Hospital Nijmegen St. Radboud, The Netherlands

J.P.L. Brokx, PhD
Institute for the Deaf, Sint Michielsgestel, The Netherlands

J.P. Donnelly, PhD
Department of Haematology, University Hospital Nijmegen St. Radboud, The Netherlands

J.B.J.M. Jansen MD PhD
Department of Gastroenterology, University Hospital Nijmegen St. Radboud,
The Netherlands

E.M. Konst, MSc
Department of Orthodontics and Oral Biology, University of Nijmegen, The Netherlands

R.J.F. Laheij, PhD
Department of Gastroenterology, University Hospital Nijmegen St. Radboud,
The Netherlands

E.H. van de Lisdonk, MD PhD
Department of General Practice, University of Nijmegen, The Netherlands

J.F.G.M. Meis, MD PhD
Department of Medical Microbiology, University Hospital Nijmegen St. Radboud,
The Netherlands

J. Mulder
Department of Medical Statistics, University of Nijmegen, The Netherlands

B.E. de Pauw, MD PhD
Department of Haematology, University Hospital Nijmegen St. Radboud, The Netherlands

M.S. van Roosmalen, MSc
Department of Medical Technology Assessment, University of Nijmegen, The Netherlands

G.S. Sonke, MSc
Department of Epidemiology, University of Nijmegen, The Netherlands, and
Department of Urology, University Hospital Nijmegen St. Radboud, The Netherlands

A.L.M. Verbeek, MD PhD
Department of Epidemiology, University of Nijmegen, The Netherlands

P.E. Verweij, MD PhD
Department of Medical Microbiology, University Hospital Nijmegen St. Radboud,
The Netherlands

P.F. de Vries Robbé, MD PhD
Department of Medical Informatics, University of Nijmegen, The Netherlands

G.J. van der Wilt, PhD
Department of Medical Technology Assessment, University of Nijmegen, The Netherlands

DANKWOORD

Uit de lijst van co-auteurs mag duidelijk zijn dat deze proeve van bekwaamheid tot stand kwam dankzij de medewerking van veel mensen. Ik wil mijn co-auteurs dan ook bedanken voor de prettige samenwerking in de afgelopen jaren en ik hoop dat nog veel interessante projecten volgen.

Beste Pieter de Vries Robbé, vanaf de start van de MTA-afdeling in Nijmegen heb jij je ingezet voor het wel en wee van de afdeling en van mijn proefschrift. Alhoewel jouw directe betrokkenheid door de verzelfstandiging van de afdeling de laatste tijd vanzelfsprekend minder is geworden, was jouw creatieve inbreng en enthousiasme bij zowel afdeling als promotieonderzoek zeer waardevol. Beste Frans Rutten, sinds ik begon als student-assistent bij jouw promovendi in Maastricht is het mij duidelijk dat jij een goed overzicht hebt van de wereld van het MTA-onderzoek. In de afrondingsfase van het proefschrift heb je me ervan kunnen overtuigen om verschillende onderdelen te herschikken om zo de puntjes op de i te zetten. Beste Gert-Jan van der Wilt, jouw rol is doorslaggevend geweest. Op die ene dag in februari dat we er eens de tijd voor namen, bedacht jij uit materiaal dat grotendeels voorhanden was een structuur en een theoretisch kader voor het werk. Jouw idee dat dit dan voor de jaarwisseling zou moeten resulteren in het manuscript leek mij zeer ambitieus, maar je uitdaging kwam op het juiste moment. Ongeacht je hoge werklast bleef je daarna tijd maken om voor mij als klankbord te dienen; een bewijs van een hoge gestandaardiseerde sukkel-dichtheid op onze werkkamer? Dank voor jullie onmisbare hulp bij de totstandkoming van dit proefschrift.

Gegevens en verwerking van gegevens zijn essentieel geweest voor bijna alle onderdelen van mijn onderzoek. Ik wil dan ook de mensen bedanken van wie en met wie ik allerlei gegevens en informatie heb kunnen gebruiken voor het uitwerken van mijn ideeën. Beste André Verbeek, jij gaf me de dataset op basis waarvan het model over de testsequenties werd ontwikkeld. Je constructieve suggesties hebben een belangrijke bijdrage geleverd aan dit onderdeel. Beste Gabe Sonke, op jouw dataset heb ik het testsequentie-model fraai kunnen toepassen. Beste Robert Laheij, niet alleen de diagnostiek dataset, maar vooral ook de arbeidsverzuimgegevens die jij tijdens je dyspepsie-projecten verzamelde bleken waardevol. Beste Paul Verweij, Jacques Meis en Peter Donnelly, het modelleringswerk op het terrein van infecties steunt op jullie gegevens en medisch-inhoudelijke kennis en ervaring. Beste Emmy Konst, de cijfers uit het PSOT-project waren een prima basis voor het proberen van nieuwe statistische methoden.

En natuurlijk zijn datasets niks zonder goed datamanagement: beste Wim en Liesbeth Lemmens, Albert Reintjes, Jan Mulder en Leo van Rossum, dank voor jullie bereidheid om

raad en daad te stellen bij dataverwerking en om steeds weer additionele analyses te doen. En beste Theo de Boo, zonder jouw statistische adviezen was ik gegarandeerd vaker in allerlei valkuilen gelopen. Daarnaast zijn om alles draaiende te houden in dit soort werk kom-sputters onontbeerlijk. Beste Rob Reuzel, zeker in de afrondingsfase van dit proefschrift was jouw onmisbare hulp geruststellend, vooral tijdens mijn ruzie met software en printers (zie omslag). Beste Hans Groenewoud en Pieter Zanstra, voor mij als adactylognost (Harry Severens, De Volkskrant d.d. 17 mei 1997) is technische ondersteuning essentieel, ook op zondagmorgen! Hand- en spandiensten werden verleend door student-assistenten, zowel voor het proefschrift zelf als voor andere ontlastende werkzaamheden: Jolanda Habraken, Leandra de Winter, Kirsten Gertsen en Joep Duijnstee, bedankt voor jullie enthousiaste hulp. Beth White, jou wil ik bedanken voor het corrigeren van mijn Engels, zelfs nadat je terug was gegaan naar de States bleef je bereid via email voor me te werken.

Ook de andere collegae van de afdeling MTA, Paul Krabbe, Margriet Hartman, Patricia Lottman, Mieke Nieuwenhuizen en Gina Wielink wil ik bedanken voor de goede samenwerking en bereidheid klussen voor elkaar op te knappen. Zeker in een multidisciplinaire groep als de onze zijn ruggespraak, referaten en seminars een prima bron voor reflectie en inspiratie. Collegae en ex-collegae van de andere afdelingen van MIES - Medische Informatiekunde, Epidemiologie en Statistiek: ik vind het belangrijk dat er naast werken (weer) tijd is voor gezellige dingen: koffie met vlaai, de kroeg in tijdens de LVVDM-VGB, het bos in om wat hard te lopen, en op de racefiets genieten van de omgeving van Nijmegen.

Verder bedank ik mijn vrienden die me stimuleerden om ook wat anders te doen dan achter mijn bureau te zitten: Les Cochons en vrienden-speleologen, bedankt dat jullie me regelmatig eens ouderwets onder de grond vergezelden; de Hardlopers In Training oftewel HIT-ters, dank voor jullie gezelschap bij de snelle kilometers en lange duurlopen als afleiding van al het denk- en schrijfwerk: mijn gezeur en geklaag zal nu wel voorbij zijn.

En Anita, jij.....

LIST OF PUBLICATIONS

First author

Severens JL, Boo ThM de & Konst EM (1999). Uncertainty of incremental cost-effectiveness ratios: a comparison of Fieller and bootstrap confidence intervals. *International Journal of Technology Assessment in Health Care* 15: 608-614.

Severens JL, Boo ThM de, Roosmalen MS van, Verweij PE & Wilt GJ van der. Validity of willingsness-to-pay for non-decisional diagnostic information [submitted].

Severens JL, Brokx JPL & Broek P van den (1997). Cost analysis of cochlear implants in deaf children in the Netherlands. *American Journal of Otology* 18: 714-718.

Severens JL, Donnelly JP, Meis JFGM, Vries Robbé PF de, Pauw BE de & Verweij PE (1997). Two strategies for managing invasive aspergillosis: a decision analysis. *Clinical Infectious Disease* 25: 1148-1154.

Severens JL, Konst EM & Prahl C (1998). Principes van kosten-effectiviteitsanalyse bij stem-spraak- en taalstoornissen. *Stem-, Spraak- en Taalpathologie* 7: 158-167.

Severens JL, Laheij RJF, Jansen JBMJ, Lisdonk EH van de & Verbeek ALM (1998). Estimating the cost of lost productivity in dyspepsia. *Alimentary Pharmacolology & Therapeutics* 12: 919-923.

Severens JL, Mulder J, Laheij RJF & Verbeek ALM. Precision and accuracy in measuring absence from work as a basis for calculating productivity costs. *Social Science & Medicine* [accepted for publication].

Severens JL, Oerlemans HM, Weegels AJPG, Hof MA van 't, Oostendorp RAB & Goris RJA (1999). Cost-effectiveness analysis of adjuvant physical or occupational therapy for patients with reflex sympathetic dystrophy. *Archives of Physical Medicine and Rehabilitation* 80: 1038-1043.

Severens JL, Prahl C, Kuijpers-Jagtman AM & Prahl-Andersen B (1998). Short-term cost-effectiveness analysis of presurgical orthopedic treatment in children with complete unilateral cleft lip and palate. *Cleft Palate - Craniofacial Journal* 35: 222-226.

Severens JL. Sonke GS, Laheij RJF, Verbeek ALM & Vries Robbé PF de. Efficient diagnostic test sequence: applications of the probability modifying plot [submitted].

Severens JL, Verweij PE, Bos JJ, Donnelly JP & Meis JFGM. Cost-effectiveness of liposomal amphotericin B for the treatment of invasive fungal infections in neutropenic patients: a decision analysis [submitted].

Severens JL, Vries Robbé PF de & Verbeek ALM (1999). Optimizing diagnostic test sequences: the probability modifying plot. *Methods of Information in Medicine* 38: 50-55.

Severens JL & Wilt GJ van der (1999). Economic evaluation of diagnostic tests: a review of published studies. *International Journal of Technology Assessment in Health Care* 15: 480-496.

Co-author

Albers JMC, Kuper HH, Riel PLCM van, Prevoo MLL, Hof MA van 't, Gestel AM van, & Severens JL (1999). Socioeconomic consequences of rheumatoid arthritis in the first year of the disease. *British Journal of Rheumatology* 38: 423-430.

Baltussen RMPM, Wielink G, Stoevelaar HJ, Wilt GJ van der, Severens JL & Ament AJHA (1998). The economic impact of the introduction of TUMT in the treatment of BPH: a scenario analysis. *World Journal of Urology* 16: 142-147.

Braspenning JCC, Severens JL, Brokx JPL & Broek P van den. Cochlear Implants for children: quality of life and cost [submitted].

Jager GJ, Severens JL, Thornbury JR, Oosterhof GON, Ruijs JHJ & Barentsz JO. Is the utilization of MR imaging in local staging of prostate cancer appropriate? *Radiology* [accepted for publication].

Laheij RJF, Jansen JBMJ, Lisdonk EH van de, Severens JL & Verbeek ALM (1996). Symptom improvement through eradication of Helicobacter pylori in patients with non-ulcer dyspepsia. *Alimentary Pharmacololgy & Therapeutics* 10: 843-850.

Laheij RJF, Jansen JBMJ, Lisdonk EH van de, Severens JL & Verbeek ALM (1999). The prognostic value of gastrointestinal morbidity for gastric cancer. *Family Practice* 16: 129-132.

Laheij RJF & Severens JL (1997). Cost of endoscopy in economic evaluation (letter). *Gastroenterology* 113: 223-224.

Laheij RJF, Severens JL, Jansen JBMJ, Lisdonk EH van de & Verbeek ALM (1997). Management in general practice of patients with persistent dyspepsia; a decision analysis. *Journal of Clinical Gastroenterology* 25: 563-567.

Laheij RJF, Severens JL, Lisdonk EH van de, Verbeek ALM & Jansen JBMJ (1998). Randomised controlled trial of omeprazole or endoscopy in patients with persistent dyspepsia; a cost-effectiveness analysis. *Alimentary Pharmacololgy & Therapeutics* 12: 1249-1256.

Laheij RJF, Severens JL, Verbeek ALM & Jansen JBMJ (1998). Cost effectiveness of treatment for gastrooesophageal reflux disease (letter). *Gut* 43: 728-729.

Oerlemans HM, Oostendorp RAB, Boo ThM de, Laan L van der, Severens JL & Goris RJA. Randomised controlled clinical trial of adjuvant physiotherapy versus occupational therapy in patients with reflex sympathetic dystrophy / complex regional pain syndrome I *Archives of Physical Medicine and Rehabilitation* [accepted for publication].

Roosmalen MS van, Severens JL, Meis JFGM, Lees E, Barton R & Verweij PE (1998). Prevalence of antibodies to Histoplasma capsulatum among Dutch speleologists (letter). *Journal of Infection* 37: 200-201.

CURRICULUM VITAE

Hans Severens werd geboren op 9 juli 1963 te Voerendaal. In 1982 haalde hij zijn Atheneum A diploma, waarna hij gedurende 3 jaar aan de Academie voor Toegepaste Kunst in Maastricht de opleiding tot edelsmid volgde. In 1985 werd gestart met de studie Gezondheidswetenschappen aan de toenmalige Rijksuniversiteit Limburg en hij studeerde in 1989 af in de richting Beleid en Beheer van de Gezondheidszorgvoorzieningen. Gedurende de laatste 3 studiejaren werkte hij als student-assistent bij de capaciteitsgroep Economie van de Gezondheidzorg (Prof. dr. E. van Doorslaer en Prof. dr. R. Janssen). Na afstuderen werd hij als toegevoegd onderzoeker aangesteld bij het instituut Medical Technology Assessment (Prof. dr. F. Rutten) van deze universiteit. In 1990 en 1991 werkte hij in Arnhem als adviseur gezondheidszorg bij OHRA verzekeringen, afdeling KLOZ/KPZ-regiovertegenwoordiging. Voor een periode van een jaar vertrok hij naar Afrika en werkte als logisticus voor Artsen Zonder Grenzen in noordelijk Ethiopië, Zuid Soedan en Addis Abeba (Ethiopië). Sinds november 1992 werkt hij als wetenschappelijk onderzoeker bij de toen in samenwerking met het Academisch Ziekenhuis Nijmegen St. Radboud opgerichte afdeling Medical Technology Assessment aan de Katholieke Universiteit Nijmegen (Dr. G. van der Wilt). Vanaf januari 1999 is sprake van een deeltijds detachering bij de Werkgroep Onderzoek Kwaliteit (WOK) aan deze universiteit (Prof. dr. R. Grol). In 1999 verbleef hij enkele maanden als visiting fellow aan de University of York, Engeland, en deed onderzoek bij het Centre for Health Economics (Prof. dr. M. Drummond, Dr. M. Sculpher en L. Fenwick) en het Department for Economics and Related Studies (Dr. K. Claxton). Recentelijk won hij van de Nederlandse Vereniging voor Technology Assessment in de Gezondheidszorg (NVTAG) de MTA-Prijs 1999.

Stellingen

behorende bij het proefschrift

Some Methodological Issues in Economic Evaluation in Health Care

1. Het rangschikken van medische technologieën naar kosten-effectiviteit is onverantwoord *(dit proefschrift)*.

2. Aanvragers van diagnostische tests vragen zich onvoldoende af of die aankoop zijn geld waard is *(dit proefschrift)*.

3. Modellering is een oplossing voor de wrijving tussen de informatiebehoefte van beleidsmakers en die van onderzoekers *(dit proefschrift)*.

4. De huidige methoden voor de berekening van productiviteitskosten leiden tot een overschatting van deze kosten *(dit proefschrift)*.

5. De rol van statistische methoden in economische evaluaties is om schijnzekerheid te vermijden *(dit proefschrift)*.

6. Kosten-effectiviteitsratio's dwingen opdrachtgevers en gebruikers van economische evaluaties tot het vaststellen van referentiekaders *(dit proefschrift)*.

7. Speleologen zijn niet rationeel *(dit proefschrift)*.

8. De indeling van kostensoorten naar direct en indirect, en medisch en niet-medisch is onnodig verwarrend.

9. De term 'kosten-baten analyse' wordt doorgaans foutief gehanteerd.

10. Evaluaties van medische technologieën dienen niet alleen de relatieve doelmatigheid, maar ook de financiële consequenties te bepalen.

11. Het nieuwe voorgestelde medicijnbeleid (commissie-De Vries) zal niet leiden tot een toename van de concurrentie tussen zorgverzekeraars.

12. De functie van het publiek bij het promotieritueel is het bijdragen aan de stress en het gevoel van succes van de promovendus (Bram Buunk, De Volkskrant d.d. 20 maart 1999).

13. What good's a disease that won't hurt you? (Lou Reed, The Thesis).

14. Hardlopers zijn doorlopers.

J.L. Severens
Nijmegen, 2 december 1999